Maestría en Bioinformática y Biología Computacional

Caracterización de regiones genómicas asociadas a rasgos cuantitativos de producción, altura de la planta y tamaño del grano en *Coffea arabica* L.

NARMER FERNANDO GALEANO VANEGAS

UNIVERSIDAD CATÓLICA DE MANIZALES
MAESTRÍA EN BIOINFORMÁTICA Y BIOLOGÍA COMPUTACIONAL
MANIZALES
2019

Caracterización de regiones genómicas asociadas a rasgos cuantitativos de producción, altura de la planta y tamaño del grano en *Coffea arabica* L.

NARMER FERNANDO GALEANO VANEGAS

Trabajo de grado presentado como requisito para optar al título de:
Magister en Bioinformática y Biología Computacional

Director:
CARLOS ERNESTO MALDONADO LONDOÑO

Co - directora:
GLORIA MARÍA RESTREPO FRANCO

UNIVERSIDAD CATÓLICA DE MANIZALES
MAESTRÍA EN BIOINFORMÁTICA Y BIOLOGÍA COMPUTACIONAL
MANIZALES
2019

Acta de calificación

Caracterización de regiones genómicas asociadas a rasgos cuantitativos de producción, altura de la planta y tamaño del grano en *Coffea arabica* L.

NARMER FERNANDO GALEANO VANEGAS

Tesis para optar el título de:
Magister en Bioinformática y Biología Computacional

Aceptación Jurado:

_____

Jurado 1

_____

Jurado 2

Tabla de contenido

Agradecimientos 3
Resumen 5
Abstract 6
Lista de tablas 7
Lista de figuras 8
Lista de anexos 9
1. Introducción 10
1.1. Campo del conocimiento 11
1.2. Planteamiento del problema 13
1.3. Justificación 14
1.4. Objetivos 14
1.4.1. Objetivo general 14
1.4.2. Objetivos específicos 15
2. Referente teórico y antecedentes 16
2.1. Genoma 16
2.1.1. Genómica de las plantas cultivadas 17
2.1.2. Secuenciación 22
2.1.3. Relación genotipo y fenotipo 31
2.2. Café 35
2.2.1. Genoma del café 35
2.3. Mapeo Genético y QTLs en café. 37
2.4. Antecedentes 42
3. Materiales y métodos 46
3.1. Secuencias de genoma y transcriptoma usadas en el trabajo 46
3.1.1. Targeted sequencing BAC por BAC Secuencias PacBio® 46
3.1.2. Secuenciación del genoma completo 46
3.1.3. Librerías de secuencias completas de Genes (Iso-seq™ – PacBio®) 46
3.2. Identificación de las secuencias de regiones que contienen QTLs de importancia agronómica en café. 49
3.2.1. Delimitación de las secuencias de regiones de interés del ensamblaje WGS 49
3.2.2. Ensamblaje de datos targeted sequencing de regiones QTLs de Tamaño del grano, producción y altura de la planta. 49
3.3. Caracterización de los genes contenidos en QTLs de importancia agronómica en café, con énfasis en los genes predichos y expresados. 50
3.3.1. Anotación de genes codificantes 50
3.3.2. Identificación de transposones, regiones repetitivas y genes relacionados con elementos transponibles 51
3.3.3. Genómica comparativa 51
4. Resultados y discusión 52
4.1. Delimitación de las regiones de interés 52

4.2. Ensamblaje de datos *targeted sequencing* de las regiones QTL de tamaño del grano, producción y altura de la planta 55
4.2.1. Ensamblaje BAC por BAC. 57
4.3. Anotación de los genes contenidos en QTLs de importancia agronómica en café. ……...64
4.3.1. Anotación de genes codificantes 64
4.3.2. Identificación de transposones, regiones repetitivas y genes relacionados con elementos transponibles..... 76
4.3.3. Genómica comparativa 78
5. Conclusiones generales 80
5.1. Contribuciones de la Tesis 80
5.2. Impactos Potenciales de la Tesis..... 80
5.3. Recomendaciones y trabajos futuros 80
6. Referencias bibliográficas 82
ANEXOS ……90

Agradecimientos

Instituciones y oficinas

Agradezco al Centro Nacional de Investigaciones de Café-Cenicafé, por darme acceso a los datos y permitirme hacer esta Tesis de Maestría. Al AGI de la Universidad de Arizona por el diseño y producción de datos. A la Universidad Católica de Manizales-UCM por su apoyo para la realización de la Maestría en Bioinformática y Biología Computacional. Al Instituto de Investigación en Microbiología y Biotecnología Agroindustrial-IMBA de la UCM, por su soporte en horas e infraestructura computacional para la realización de la presente Tesis. Al Grupo de Mejoramiento Genético de Cenicafé y al Grupo de Investigaciones Biológicas-GIBI, por el apoyo permanente durante el desarrollo de la investigación.

Doctores e investigadores

- Al Dr. Carlos Ernesto Maldonado Londoño. Investigador de Cenicafé, Director de la Tesis, por permitirme trabajar con datos fruto de su diseño experimental y trabajo de laboratorio. Por su mirada crítica y orientación en el desarrollo de la misma, que me llevó a adquirir conocimientos y aclarar conceptos que fueron muy útiles en su desarrollo.
- A la Dra. Gloria María Restrepo Franco, Codirectora de la Tesis, por su paciencia, por enseñarme a perseverar y compartir sus conocimientos conmigo para poder llevar a cabo este trabajo de grado y terminar el presente documento. Además, debo agradecer su amistad que me permitió no desfallecer a pesar de las circunstancias.
- Al Dr Rod A Wing por el acompañamiento en diseño del experimento
- A la Dra Pilar Moncada por el suministro del material vegetal para secuenciación de Iso-seq e Illumina
- Al Dr. Alvaro Gaitan quien suministró tejido para secuenciación genómica PacBio y por su apoyo a esta investigación como director de Cenicafé.

Otras personas

- A Beatriz Padilla por su trabajo en el laboratorio para la generación de librerías y la obtención de los datos usados para la realización de esta tesis.
- Agradezco infinitamente a todos mis compañeros del IMBA, los actuales y los pasados, porque sin ellos y su apoyo incondicional hubiese sido imposible poder concretar esta Tesis.
- A mis compañeros y amigos de la Maestría, porque gracias a la interacción que tuvimos, aprendí muchas cosas a nivel técnico y personal que me ayudaron en llevar a feliz término este trabajo a pesar de los obstáculos normales.

Agradecimientos personales

- A mi Mamá Maria Luz Vanegas y mi Papá Rodrigo Galeano, por su apoyo emocional y espiritual, que a pesar de la adversidad siempre están a mi lado, dándome todo su cariño

y amor. A mi hermana Tatiana Galeano y mi tía Luisa Vanegas, por su cariño y por estar ahí cuando más las necesito.

- A mi señora, Luisa Vásquez, porque tuvo que soportar con cariño mis cambios de humor y de ánimo en el transcurrir de este trabajo de grado, y me dio mucho café para poder pasar las noches en vela y poder llegar al final de este proceso, sin ella tal vez me hubiera rendido, pero su fortaleza me ayudó a llevar la adversidad con su ejemplo.

Resumen

El café es la segunda materia prima más comercializada en el mundo y es la fuente de ingresos de más de 125 millones de personas. Dentro de las especies cultivadas, *Coffea arabica* y *C. canephora* lideran el área plantada en todo el mundo. El género *Coffea* se compone de aproximadamente 100 especies con un punto de radiación calculado de 0,5 millones de años, siendo *C. arabica* una de las especies más recientes (~ 0,1 millones de años) y el único tetraploide del género. Su genoma fue el resultado de un híbrido fértil entre *C. canephora* y *C. eugenioides*. Los subgéneros de *C. arabica* tienen una alta identidad entre ellos, lo que ha obstaculizado el proceso de ensamblaje del genoma. De acuerdo con lo anterior, el objetivo del presente trabajo fue caracterizar la secuencia de regiones genómicas asociadas a rasgos cuantitativos de producción, altura de la planta y tamaño del grano en *Coffea arabica.*

La caracterización de una región genómica que contenía un QTL para rendimiento, en términos de variantes estructurales, sintenia y contenido de genes, se comparó a través de dos enfoques de secuenciación; lo anterior con el fin de identificar la estrategia más eficiente para obtener el mejor ensamblaje de la región blanco, de un genoma allotetraploide.

Se implementó un enfoque de secuenciación dirigida, para caracterizar regiones genómicas que contenían QTL asociados con rasgos de agronomía importantes. El MTP que cubre las regiones fue seleccionada, secuenciada por PacBio y ensamblada por los *software* HGAP y postHGAP. Un segundo conjunto de datos de todo el genoma fue obtenido por SMRT-seq a partir de bibliotecas de 20 Kb (WGS-SMRT, PacBio) y ensamblado usando Falcon. Del transcriptoma de los tejidos (meristemos, hojas y flores), se obtuvieron transcritos por secuenciación de cDNA de longitud completa (Iso-Seq PacBio) de *C. arabica* var. Caturra. Para las dos estrategias de ensamblaje se tomó una región que abarcó 30 cM en direcciónes 5’ y 3’ de los marcadores de los marcadores (SNV, DArT y SSR) asociados con el QTL. El *C. arabica* WGS se alineó con *C. canephora* y, como se esperaba, varios *contigs* WGS mapearon sobre la misma región de referencia; se asignaron *contigs* con mayor identidad a *C. canephora* (C^a) a ese subgénero y a aquellos con menor identidad asignada al subgenoma de *C. eugenioides* (E^a). Los *contigs* de *contigs* que no se superponen se asignaron a ambos subgéneros en el proceso de *scaffolding*. En el caso de la secuenciación BAC-by-BAC, el ensamblaje se realizó utilizando Minimus. La anotación de predicciones y repeticiones de genes se realizó utilizando el programa MAKER, y la anotación funcional se realizó con InterProscan.

Se obtuvo una secuencia ensamblada de 7,38 Mb con 1188 genes predichos a partir del enfoque BAC por BAC. De la estrategia WGS se obtuvieron *scaffolds* de 10,2 Mb con 1675 genes pronosticados para el subgenoma de C^a y 10,7 Mb con 1991 para subgenoma de *C. eugenioides* E^a. El análisis Dot-Plot mostró evidencia de ensamblaje quimérico entre subgéneros en el enfoque BAC por BAC. con la resistencia a enfermedades que podrían estar contribuyendo al rendimiento del café.

En el presente trabajo, la mejor estrategia de secuenciación para obtener datos de alta calidad del genoma tetraploide fue la secuenciación WGS, demostrada en ausencia de regiones quiméricas y una clara diferenciación de algunas subregiones entre subgéneros, así como la identificación de genes con variantes que podrían ser usados como marcadores moleculares para mapeo fino.

Palabras clave: *Coffea arabica, Coffea canephora, Coffea eugenioides, Targeted Sequencing, SMRT-seq,* Mapa físico, *Isoformas*

Abstract

Coffee is the second most traded raw material in the world being the source of income of more than 125 million people. Within the cultivated species *Coffea arabica* and *C. canephora* lead the planted area worldwide. The genus Coffea is made up of about 100 species with a calculated radiation point of 0.5 million years, with *C. arabica* being one of the most recent species (~ 0.1 million years old) and the only tetraploid of the genus. Its genome was the result of a fertile hybrid between *C. canephora* and *C. eugenioides*. The subgenomes of *C. arabica* have a high identity between them, which has hampered the process of genome assembly. In order to characterize a genomic region containing a QTL for yield, in terms of structural variants, synteny and gene content and collinearity, two sequencing approaches were compared to identify the most efficient strategy to obtain the better assembly of the target region of an allotetraploid genome.

A tergeted sequencing approach was implemented to characterize genetic regions that contain QTL associated with important agronomy traits. The minimum tiling path (MTP) covering the regions of interest was selected, sequenced by PacBio and assembled by the HGAP and postHGAP software. A second set of genome-wide data was obtained by SMRT-seq from 20 Kb libraries (WGS-SMRT, PacBio) and assembled using Falcon. From the transcriptome of the tissues (meristems, leaves and flowers), transcripts were obtained by sequencing full length cDNA (Iso-Seq PacBio) of *C. arabica* var. Caturra for the two assembly strategies a region covering 30 cM upstream and downstream of the markers (SNV, DArT and SSR) associated with the QTL was taken. *C. arabica* WGS was aligned with *C. canephora* and, as expected, several WGS contigs mapped over the same reference region; contigs with greater identity were assigned to *C. canephora* ($C^{a)}$ to that subgenus and to those with less identity assigned to the subgenome of *C. eugenioides* (E^{a}). The contigs of contigs that do not overlap were assigned to both subgenres in the scaffolding process. In the case of BAC-by-BAC sequencing, the assembly was performed using Minimus. The annotation of predictions and gene repetitions was made using the MAKER program, and the functional annotation was made with InterProscan.

An assembled sequence of 7.38 Mb with 1188 predicted genes was obtained from the BAC-by-BAC approach. WGS derived scaffolds of 10.2 Mb with 1675 predicted genes for the C^{a} and 10.7 Mb with 1991 for E^{a}. Dot-Plot analysis showed evidence of chimeric assembly between subgenomes in the BAC-by-BAC approach.

In the present work, the best sequencing strategy to obtain high quality data from the tetraploid genome was the WGS sequencing, demonstrated in no apparent presence of chimeric regions and a clear differentiation of some subregions between subgenomes. As the identification of genes with variants that can be used as molecular markers for fine mapping as well.

Keywords: *Coffea arabica, Coffea canephora, Coffea eugenioides,* Targeted sequencing, *SMRT-seq,* Physical mapping

Lista de tablas

Tabla 1. Resumen de secuencias que fueron usadas en el estudio. 46
Tabla 2. Secuencias pertenecientes a las diferentes librerías en los parentales de la población del mapa genético de café. 47
Tabla 3. Tamaño de Regiones de interés en *C. arabica* y *C. canephora WGS*. 53
Tabla 4. Organización de los *contigs* de las Regiónes de interés y asignación a los subgenomas. 55
Tabla 5. Ensamblaje BAC por BAC de la región YLD2 y sus estadísticas. 58
Tabla 6. Ensamblaje BAC por BAC de la región PH5 y sus estadísticas. 60
Tabla 7. Genes dentro de los rangos de marcadores moleculares mapeados por región 70

Lista de figuras

Figura 1. La poliploidía y su efecto en el genoma de las plantas. 18
Figura 2. Ensamblaje del genoma y *phasing* de haplotipos.. 28
Figura 3. Plasticidad fenotípica e interacción genotipo ambiente (G × A).. 32
Figura 4. El papel de la fisiología en la interacción (G × A × M). 34
Figura 5. QTLs identificados en la población F3 para producción (*yield*), altura de la planta (*plant height*) y tamaño del grano (*bean size*). 43
Figura 6. Flujo de trabajo para el análisis de las regiones QTL asociadas a producción, tamaño del grano y altura de la planta. 48
Figura 7. Ejemplo delimitación regiones de acuerdo a los marcadores moleculares asociados a producción, altura de la planta y tamaño del grano. 52
Figura 8. Delimitación de las regiones de interés en los grupos de ligamiento para los marcadores asociados a producción, tamaño del grano y altura de la planta. 53
Figura 9. Mapeo de *contigs* WGS de la región YLD2 contra el cromosoma 5 de *C. canephora*. (A). Salida del SyMAP. (B). Subregión del rasgo YLD2 y asignación de *contigs* a los subgenomas de C. arabica de acuerdo a la identidad, mayor identidad (Rojo) asignada a C^a y Menor identidad (Rojo claro con amarillo) es asignado a E^a. 55
Figura 10. D*ot plots* de los subgenomas C^a (Eje Horizontal) Vs. E^a (Eje Vertical). 57
Figura 11. Minimum Tiling Path. A. Región Asociada a producción YLD2. B. MTP visualizado en el Mapa Físico. 58
Figura 12. Dot plot de Región YLD2 en *C. canephora* contra los Contigs FPC pertenecientes al MTP. 59
Figura 13. Dotplot de las comparaciones entre los scaffolds de los subgenomas y el ensamblaje BAC por BAC de la región YLD2.. 59
Figura 14. Dot plot de Región PH5 en *C. canephora* contra los Contigs FPC pertenecientes al MTP. 61
Figura 15. Dotplot de las comparaciones entre los scaffolds de los subgenomas WGS (Eje Horizontal) y el ensamblaje BAC por BAC (Eje Vertical) de la región PH5. 62
Figura 16. Elementos transponibles para el tamaño del Grano.. 76
Figura 17. Elementos transponibles para las regiones asociadas a altura de la planta.. 77
Figura 18. Elementos transponibles para las regiones asociadas a producción. 78
Figura 19. Anotación del PFAM para los subgenomas de *C. canephora* y C. *eugenioides.* 69

Lista de anexos

Anexo I. Comparación de las secuencias de los subgenomas de los distintos rasgos con su región correspondiente en *C. canephora* 90
Anexo II. Comparación de las secuencias de los subgenomas de los distintos rasgos..... 93
Anexo III. Anotación de los genes en las regiones de interés según bases de datos PFAM. 95

Tabla S 1. Genes Región BS4eug. 95
Tabla S 2. Genes Región PH1can. 97
Tabla S 3. Genes Región PH1eug. 100
Tabla S 4. Genes Región PH4eug. 103
Tabla S 5. Genes Región PH5can. 105
Tabla S 6. Genes Región PH5eug. 112
Tabla S 7. Genes Región YLD2can. 122
Tabla S 8. Genes Región YLD2eug. 130

1. Introducción

El café es el 117º producto agrícola más comercializado en el mundo (OEC MIT, 2017). Su cultivo es de gran importancia para la economía colombiana y de 69 países más, que producen 9 millones de toneladas al año, con un valor total del producto exportado de USD 24.000 millones de dólares para el 2012 (FAO, 2015). Colombia se encuentra en el tercer lugar en la lista de exportadores de café según registros de la FAO con una producción anual de 745,084 t de café para el año 2016 (FAO, 2018). La producción nacional en el año 2015 fue de 12´713.667 sacos de 60 kg, con un valor de exportación de USD 2.788´362.488 dólares según datos de la Federación Nacional de Cafeteros (2017). De este cultivo dependen aproximadamente 500.000 familias cafeteras (FNC, 2018).

El género *Coffea*, está conformado por alrededor de 100 especies, con una historia evolutiva muy reciente (Hamon *et al*., 2017), ya que se calcula un punto de radiación de apenas 0,1 a 0,5 millones de años (BP) (Anthony, Diniz, Combes, & Lashermes, 2010), siendo *Coffea arabica* una de las especies más recientes (~0,1 millones de años) (Cenci, Combes, & Lashermes, 2012) y el único tetraploide del género (Lashermes, Hueber, Combes, Severac, & Dereeper, 2016). Dentro de las especies cultivadas se encuentran *C. arabica, C. canephora* y *C. liberica*, siendo las dos primeras las de mayor importancia en el mercado mundial (Fain, Quiñones, Álvarez-Berríos, Parés-Ramos, & Gould, 2017).

Aunque *C. arabica* es considerado como el café con mayor calidad de taza, continúa siendo una de las prioridades en investigación el mejoramiento de la calidad de las dos especies comerciales (*C. arabica* y *C. canephora*). Con los avances en genómica y la tecnología de secuenciación, es factible entender el genoma del café y asociarlo con rasgos de importancia como la calidad, la producción, la morfología de planta, entre otros, así como mejorar la eficiencia de los programas de mejoramiento genético (Tran, Lee, Furtado, Smyth, & Henry, 2016). El diseño para la obtención de nuevas variedades por el programa de mejoramiento genético en Cenicafé, tiene una duración de 20 a 35 años, dados los largos ciclos de cultivo y selección (Rivera Calderón, 2015). Es en este escenario, la selección asistida por marcadores moleculares se convierte en un factor fundamental para la reducción de los tiempos en los programas de mejoramiento, que para *C. arabica* puede representar una reducción de al menos 10 años en la obtención de una nueva variedad (Comunicación personal Maldonado, 2017).

En *C. canephora* se cuenta con la identificación de QTLs que controlan caracteres agronómicos relacionados con calidad, con un mapa genético de 236 marcadores. Para la descripción de los QTLs se ha evaluado la calidad de la bebida, relacionando los rasgos bioquímicos y de degustación de taza. En los QTLs se han identificado casi la mitad de los caracteres evaluados, con efectos que van desde el 6% al 80 % de la variación fenotípica. La mayoría de ellos presentan una detección consistente durante años. Los QTLs más fuertes explican un alto porcentaje de la variación de rendimiento (34% a 57%), el tamaño del grano (25% a 35%), contenido de ácidos clorogénicos (22% a 35%), sacarosa y el contenido de trigonelina (29 % a 81%), y la acidez y el amargor de las bebidas de café (30% a 55%). En estas regiones han sido identificados genes que codifican para la biosíntesis de la cafeína, en la biosíntesis de los ácidos clorogénicos, y genes implicados en el metabolismo del azúcar (Leroy, De Bellis, Legnate, Kananura, Gonzales, Pereira, Andrade, Charmetant, Montagnon, Cubry, Marraccini, Pot, & de Kochko, 2011a).

Recientemente, fue publicado el mapa genético de *C. arabica* en el que se identificaron QTLs asociados con producción, altura y tamaño del grano, donde fueron descritos 22 grupos de ligamiento usando 338 marcadores para su construcción. El rasgo de producción fue asociado con un marcador el cual explicó la variación del 22,2% del 14, 35% de acuerdo con los ambientes evaluados en el grupo de ligamiento dos. En el rasgo de producción también fue asociado un QTL que explicó el 16,9% de la variación de los datos en el grupo de ligamiento 11. En el caso del rasgo asociado a la altura de la planta, se detectó un QTL en el grupo de ligamiento uno que explicó el 25 % de los datos y en el grupo de ligamiento cuatro que explicó el 15,65% de los datos. Finalmente para el rasgo de tamaño del grano sólo se describen dos QTLs detectados usando el análisis de un solo punto, uno ubicado en el grupo de ligación 4 y otro en el grupo de ligación seis (Moncada *et al*., 2016).

Paralelamente, Cenicafé desarrolló un mapa físico de *C. arabica* a partir de una librería BAC con secuencias *BAC End Sequences* (BES) que permiten integrar la secuencia genómica publicada, con el mapa genético y el mapa físico de *C. arabica*. De acuerdo con lo anterior, el objetivo del presente trabajo fue caracterizar la secuencia de regiones genómicas asociadas a rasgos cuantitativos de producción, altura de la planta y tamaño del grano en *Coffea arabica.*

1.1. Campo del conocimiento

El ensamblaje del genoma de las plantas presenta problemas como el tamaño del genoma, su contenido repetitivo, la paralogía y la heterocigosidad los cuales son comunes en estos genomas (Bolger *et al*., 2017). A medida que la producción de datos ha aumentado debido al auge de las tecnologías *Next Generation Sequencing* (NGS) y su rendimiento se ha hecho que sea económico secuenciar la mayoría de los genomas de las plantas cultivadas, pero la resolución de la complejidad del genoma de la planta con lecturas pequeñas es un desafío. La mayoría de los genomas de plantas en borrador permanecen en miles de *contigs* altamente fragmentados o en cientos de *scaffolds* con numerosos *gaps*, incluso el genoma de Arabidopsis, que es posiblemente el genoma mejor ensamblado de una planta, todavía está en 102 *contigs* con una longitud total del *gap* de al menos 185.644 pb, donde la última versión contiene 29 fragmentos mal ensamblados, presenta 117 *gaps* y secuencias repetitivas perdidas, generalmente del centrómero que representan al menos 25 Mb. (Lamesch *et al*., 2011; Michael *et al*., 2018).

El tamaño del genoma y el contenido de repeticiones, que a menudo están altamente correlacionados, presentan un problema importante para el ensamblaje del genoma de las plantas. El tamaño del genoma en las plantas varía en 4 órdenes de magnitud, desde 61 Mb (*Genlisea tuberosa*) hasta más de 150,000 Mb (*Paris japonica*) (Michael, 2014). Esta gran producción de datos de las plataformas NGS ha permitido generar los suficientes datos para secuenciar genomas grandes, haciendo que el ensamblaje se haya convertido en un problema de cómputo importante. El genoma del *Pinus taeda* es el genoma más grande ensamblado hasta la fecha (22 Gb) y utilizó un conjunto preprocesado y condensado de “súper lecturas” para reducir los recursos de cálculo necesarios para el ensamblaje (Michael & VanBuren, 2015).

Las repeticiones son un problema importante en el ensamblaje del genoma, y la resolución de estructuras repetitivas requiere de lecturas que excedan la longitud de las repeticiones de 10-20 Kb que se encuentran comúnmente en los genomas de las plantas. Los retrotransposones de repetición terminal larga (LTR por sus siglas en inglés) tipo “cortar y

pegar" son las repeticiones más frecuentes en el genoma de la planta y su proliferación produce expansión del genoma. En general, la estimación de las longitudes de las regiones repetitivas en el genoma es crucial para elegir las longitudes de las lecturas, la tecnología de secuenciación, los tamaños de las bibliotecas *mate-pair* y la cobertura. Gran parte de la Variación Estructural (VE) entre los cultivares dentro de los genomas de las plantas se debe al movimiento de los LTR, y los proyectos de re-secuenciación basados en referencias a menudo fallan o predicen las variantes de un único nucleótido (SV por sus siglas en inglés) poco confiables. El ensamblaje de *novo* de tres cepas de arroz divergentes descubrió varias megabases (Mb) de secuencias nuevas en cada cultivar, con muchos *contigs* que contienen genes expresados (Schatz *et al*., 2014).

Los problemas de duplicaciones segmentales parálogas complican el ensamblaje del genoma y dan como resultado ensamblajes incompletos y altamente fragmentados. La poliploidía es común entre las especies cultivadas y las plantas tienen grandes familias de múltiples genes con parálogos muy similares (C. Li, Lin, An, Wang, & Huang, 2018).

Las regiones parálogas y los sitios heterocigotos crean "burbujas" en el grafo de De Bruijin durante el ensamblaje del genoma donde dos o más regiones que son muy similares se ensamblan juntas, y las regiones disímiles adyacentes se ensamblan por separado y eventualmente se vuelven a ensamblar (Michael & VanBuren, 2015). Los problemas de ensamblaje derivados de la poliploidía o la heterocigosidad pueden superarse mediante el uso de especies progenitoras diploides como café "robusta" (*C. canephora*) (Denoeud *et al*., 2014) y trigo silvestre (*Triticum urartum*)(Ling *et al*., 2013), especies diploides silvestres estrechamente relacionadas (fresa silvestre (*Fragaria vesca*) (Shulaev *et al*., 2010)), líneas haploides / monoploides (cítricos (G. A. Wu *et al*., 2014), banano (D'Hont *et al*., 2012) o melocotón (*Prunus persica*) (Verde *et al*., 2013)), o un cromosoma artificial bacteriano (BAC por sus siglas en inglés) usando el método de *Targeted sequencing* BAC por BAC (maíz (*Zea mays*) (Schnable *et al*., 2009) o pera (J. Wu *et al*., 2013)).

La contaminación con ADN de organelos puede ser un problema importante en los proyectos de secuenciación de genomas. Sin embargo, los genomas nucleares de plantas también contienen numerosas secuencias derivadas de organelos que pueden tener una homología casi idéntica a los genomas de los organelos mismos, la secuenciación precisa a través de estas regiones requiere longitudes de lectura que puedan abarcar los sitios de unión de estas inserciones (Michael & VanBuren, 2015).

Para superar estos desafíos en la secuenciación del genoma de las plantas se requieren avances en la tecnología de secuenciación. Las longitudes de lectura más largas proporcionadas por secuenciadores de tercera generación como *Pacific Biosciences* (PacBio) ofrecen la posibilidad de superar las complejidades de los genomas de las plantas. El tamaño promedio de las lecturas de PacBio utilizando la química P6C4 es superior a 15 kb con una cobertura promedio de 30 requerida para una calidad de Q50. La secuencia consenso circularizada (CCS por sus siglas en inglés) de PacBio usa secuencias cortas (1-5 kb) (*reads of insert*) que se circularizan y se secuencian varias veces para producir secuencias de muy alta calidad que se pueden usar para distinguir estructuras repetitivas o parálogas muy similares. CCS también es útil para construir transcritos de alta confianza, alelo específicas, de longitud completa con sitios de empalme alternativos bien anotados. Dada una profundidad de lectura lo suficientemente alta, todas las repeticiones excepto las más largas (como los telómeros y los centrómeros) se pueden resolver, lo que posiblemente marque el comienzo de una nueva era de conjuntos de genoma de calidad de platino. (Rhoads & Au, 2015; Tilgner, Grubert, Sharon, & Snyder, 2014). Como problemas de estas tecnologías de

tercera generación incluyendo las secuencias generadas por PacBio y Oxford Nanopore Technologies (ONT) tienen un problema de generación de *indels*, los cuales deben ser corregidos por secuenciación Illumina para obtener un mejor ensamblaje.

Por estas razones, en este documento se han incluido datos de la técnica de targeted *sequencing* BAC por BAC de regiones de interés para producción, altura de la planta y tamaño del grano y la secuenciación de genoma completo, haciendo énfasis en esas regiones, con el fin de comparar y validar los resultados de las dos técnicas para identificar la estrategia más eficiente para el mejor ensamblaje de la región blanco de un genoma alotetraploide.

1.2. Planteamiento del problema

Un genoma de referencia ensamblado es un patrón de comparación para estudios de mejoramiento genético, donde el desafío es identificar las características del genoma que permiten asociar y describir la biología del organismo en estudio. En el caso del genoma, las asociaciones de los rasgos agronómicos con este se realizan mediante el uso de marcadores moleculares, los cuales sirven como conexión entre, la información del mapa genético y los rasgos identificados, con la secuencia del genoma. Una vez se generen mapas genómicos detallados, se podría adquirir la capacidad de editar secuencias específicas, como lo que se ha hecho en arroz y trigo (Shan, Wang, Li, & Gao, 2014), permitiendo la próxima generación de cultivos.

El mejoramiento genético ha sometido a los genomas de los cultivos a un cuello de botella para satisfacer las necesidades de la población. El impacto de este proceso es la pérdida de la diversidad en el germoplasma, desencadenando en un mejoramiento lento y en la potencial pérdida de características como la resistencia a las enfermedades. Por lo tanto, existe la necesidad apremiante de desarrollar recursos genómicos para que los parientes silvestres de los cultivos puedan ser utilizados en el mejoramiento, identificación de alelos e introgresión (McCouch *et al*., 2013).

En cultivos poliploides de baja heterocigosis como el café, se disponen de genomas de referencia diploides o haploides, en contraste con la mayoría de los genomas altamente heterocigóticos que tienen sólo un haplotipo secuenciado. Algunos de los elementos que contribuyen a los rasgos agronómicos de interés para el mejoramiento genético, como la duplicación de genes, el reordenamiento del genoma y la integración de repeticiones son específicas de cada cultivar, y estos no se pueden encontrar fácilmente utilizando estrategias de resecuenciación o referencias de complejidad reducida. De acuerdo con lo anterior se necesitan referencias de alta calidad de cada subgenoma (en poliploides) y de cada haplotipo (en cultivos heterocigotos), así como múltiples genomas de referencia por especie de cultivo para estudiar la variación intraespecífica (Michael & VanBuren, 2015).

La brecha del conocimiento en el caso específico de *Coffea arabica* es la falta de asociación entre la información del mapa genético de café (Moncada *et al*., 2016), donde se identifican las regiones que están relacionadas con rasgos de interés comercial como altura de la planta, tamaño del grano y producción y el mapa físico realizado en CENICAFÉ.

Esto permitirá la asociación de las regiones identificadas en el mapa genético, el cual está construido con base a distancias en términos de probabilidad o secuencias genéticas, a distancias en pb donde pueden describirse el contenido de secuencias génicas contenidas en el intervalo de confianza o QTL asociado al rasgo de interés. El *Minimum Tiling Path (*MTP) secuenciados por la tecnología PacBio junto a la secuenciación realizada *Whole Genome Sequencing* (WGS) pueden ayudar a resolver la pregunta biológica de cuáles son los genes

que están presentes en estas regiones y cuáles pueden ser responsables de la asociación de estos QTLs a estos rasgos.

Así mismo, se puede obtener la posible distribución de dichos genes en los diferentes subgenomas y poder inferir el comportamiento de estos respecto a los rasgos incluidos en el estudio.

Pregunta de Investigación

¿Cuáles son las características de la región genómica adyacente a marcadores moleculares asociados con los rasgos cuantitativos de producción, altura de la planta y tamaño del grano?

1.3. Justificación

Con los avances en genómica y tecnología de secuenciación, es factible entender el genoma del café y asociarlo con rasgos de importancia como la calidad, la producción, la morfología de planta, entre otros, así como a mejorar la eficiencia de los programas de mejoramiento genético (Tran *et al*., 2016). El diseño empleado para la obtención de nuevas variedades por el programa de mejoramiento genético en café es de 25 años, dados los largos ciclos de cultivo y selección (M. K. Mishra & Slater, 2012; Ribas, Pereira, & Vieira, 2006). Es en este escenario, donde la selección asistida por marcadores moleculares es un factor fundamental para la reducción de los tiempos en los programas de mejoramiento que para el caso de *C. arabica* puede representar una reducción de al menos 10 años para la obtención de una nueva variedad.

A pesar de la importancia económica del café, el primer borrador del genoma fue reportado para el café robusta sólo hasta el 2014 por Denoeud *et al*., pero ha demostrado ser de gran utilidad para describir los genes asociados a rasgos de interés agronómico y la evolución de estos lo cual permite entender la biología de la planta, y permite hallar nuevos candidatos para el mejoramiento asistido por marcadores, o el mejoramiento asistido por el genoma.

Partiendo de este antecedente, se decidió caracterizar las regiones genómicas que flanquean los marcadores de los QTLs asociados a producción, altura y tamaño del grano. Para esto, se contó con un mapa genético de *C. arabica* en el cual se habían identificado los rasgos antes mencionados (Moncada *et al*., 2016), y con un mapa físico de *C. arabica* a partir de una librería BAC con secuencias BES, desarrollada por Cenicafé. La disponibilidad de estos mapas permitió su integración con la secuencia genómica para describir los genes que posiblemente están involucrados con estos rasgos y poder encontrar marcadores candidatos que permitan detectar el genotipo de interés y así poder orientar el proceso de mejoramiento de la planta, para que en un futuro puedan ser usados en el desarrollo de una nueva variedad.

1.4. Objetivos

1.4.1. Objetivo general

Caracterizar la secuencia de regiones genómicas asociadas a rasgos cuantitativos de producción, altura de la planta y tamaño del grano en *Coffea arabica* L.

1.4.2. Objetivos específicos

- Identificar secuencias genómicas de regiones que contienen QTLs de importancia agronómica en café.
- Caracterizar los genes contenidos en regiones QTLs para tamaño de fruto, altura de la planta y producción.

2. Referente teórico y antecedentes

2.1. Genoma

En el siglo 20 el campo de la biología obtuvo tres logros significativos: la síntesis evolutiva moderna, el surgimiento de la biología molecular y la era ‘ómica’, las cuales aparecieron con intervalos de aproximadamente 30 años. Del último logro, una de las más importantes fue la era genómica, la cual comenzó en los años noventa. El desarrollo de las tecnologías de secuenciación permitieron que fueran secuenciados genomas completos, dando un paso más allá, donde los genomas fueron estudiados utilizando marcadores moleculares y *tags* de secuencia (Ellegren, 2014).

Poco después, el acceso a gran cantidad de datos de los genomas ya secuenciados sentó las bases para la caracterización molecular y fenotípica del genoma, acuñándose el término ‘ómica’. Actualmente, del mismo concepto se han derivado otros análisis biológicos a gran escala, los cuales tienen como objetivo estudiar la totalidad de un tipo de moléculas (p. ej., transcriptoma y proteoma), modificaciones (degradoma y metiloma) o rasgos de un organismo (p. ej., metagenoma). Esto ha marcado una tendencia cada vez más clara hacia una biología más cuantitativa e integradora, aprovechando los desarrollos tecnológicos para la generación de una gran cantidad de datos (Ellegren, 2014).

Los cultivos perennes son fuentes importantes de alimentos frescos y procesados en todo el mundo, pero los avances en el mejoramiento de estas plantas son difíciles debido a su propia naturaleza, ya que estos tardan varios años en alcanzar la madurez de producción y generalmente requieren grandes áreas para crecer, lo que hace que el mejoramiento de nuevos cultivares sea costoso en comparación con la mayoría de los cultivos anuales. Por esta razón, muchos cultivos perennes corren el riesgo de fracasar ya que, frente a la evolución de las plagas y las enfermedades, su diversidad genética es muy estrecha o nula, debido a que se cultivan en monocultivos (pequeñas cantidades de cultivares de élite que se propagan vegetativamente durante décadas o incluso siglos). Este fenómeno se da a pesar de que existe una gran diversidad genética disponible. Finalmente debido a los altos costos de la reproducción, los mejoradores de cultivos perennes tienen muchas dificultades para generar cultivares comercialmente exitosos de manera oportuna y rentable (McClure, Sawler, Gardner, Money, & Myles, 2014).

En realidad, los cultivares con mayor éxito no fueron consecuencia del mejoramiento o la reproducción direccionada, sino el resultado de "semillas aleatorias", o de plantas derivadas de fertilizaciones incontroladas; como ejemplo de lo anterior están el mango y la manzana. En el caso del mango los cultivares se obtuvieron de plantas aleatorias que mostraron buenos rendimientos, y en el caso de la manzana, la variedad McInstosh, uno de los cultivares más conocidos de Canadá, fue descubierto por un agricultor en su propiedad por accidente. Los cultivares perennes en algunos casos, ni siquiera han sido domesticados y no se pueden distinguir de sus antepasados silvestres, como lo son el cacao y algunos cultivares de banano. Esto contrasta enormemente con los cultivares anuales, donde los enormes esfuerzos de mejoramiento son la norma, y se esperan mejoras año tras año (McClure *et al*., 2014).

2.1.1. Genómica de las plantas cultivadas

Desde el año 2000 se han secuenciado 498 genomas de plantas (NCBI, 2018), el 63% de los cuales son especies cultivadas. La importancia de estas secuencias genómicas radica en la información generada sobre la arquitectura, la evolución y las características especiales de los genomas de los cultivares, como la retención de rasgos agronómicos después de los eventos de duplicación del genoma completo. Algunos cultivares tienen genomas muy grandes, poliploides, ricos en repeticiones, lo cual requiere el desarrollo y la innovación de nuevas tecnologías para la secuenciación, el ensamblaje y el análisis. Incluso los genomas de referencia de baja calidad pueden ayudar en el mejoramiento del germoplasma de los cultivares a través de marcadores moleculares de todo el genoma (MAS, Selección asistida por marcadores), lo que disminuye los costos de la fenotipificación y los ciclos de reproducción (Michael & VanBuren, 2015).

2.1.1.1. Poliploidización

La poliploidía es un fenómeno por el cual se originan células con tres o más juegos de cromosomas de la misma o diferentes especies o con dos o más genomas de múltiples especies. Esta ha contribuido sustancialmente a la diversificación de las plantas con semillas y al extraordinario éxito de las angiospermas. Entre las especies de plantas con semillas, el 35% son poliploides. Se desconoce la proporción de alopoliploides (de dos especies) y autopoliploides (de la misma especie) debido a que el recuento cromosómico es insuficiente para distinguir estos citotipos poliploides. Se cree que la formación de autopoliploides es más frecuente que la de los alopoliploides, pero estos últimos tienen más ventajas durante la fase de establecimiento debido a su heterosis; además, los autopoliploides también sufren de reducción de la fertilidad, lo cual es consistente con la observación de que hay más cultivos alopoliploides (trigo, algodón, tabaco, fresa y colza) que cultivos autopoliploides (tales como papa, caña de azúcar y banano) (Ming, 2015).

2.1.1.1.1. Clases de poliploidía

Los poliploides se clasifican en dos: autopoliploide y alopoliploide. Un autopoliploide es un individuo que surge dentro de una o entre poblaciones de una sola especie, mientras que el alopoliploide se refiere a individuos de origen híbrido (Figura 1). La identificación de cada clase se logra analizando los patrones de cruzamiento y herencia. La presencia de más de dos cromosomas homólogos en autopoliploides puede contribuir a la formación de multivalentes[1] durante la meiosis. Debido al apareamiento no preferencial, el descendiente de un cruce entre dos autopoliploides exhibirá una relación diferente a la del cruce Mendeliano clásico (1: 2: 1), caracterizando una herencia polisómica (Ramsey y Schemske 1998; Parisod 2013).

Los alopoliploides se dividen en dos subclases: alopoliploides verdaderos y segmentados. La formación de alopoliploides verdaderos implica la hibridación entre especies distantemente relacionadas. En este caso, los complementos cromosómicos divergentes no se emparejan entre sí, dando lugar a la formación de bivalentes durante la meiosis y en un patrón de herencia disómica. Por otra parte, los alopoliploides segmentados se originan de la hibridación entre especies relacionadas de forma clásica con genomas

[1] Más de un par de cromosomas homólogos pareados.

parcialmente diferenciados. Por lo tanto, los alopoliploides segmentados pueden someterse a emparejamiento univalente, bivalente y / o multivalente durante la meiosis, siendo considerados tipos intermedios entre los verdaderos alopoliploides y autopoliploides (Sattler, Carvalho, & Clarindo, 2015a).

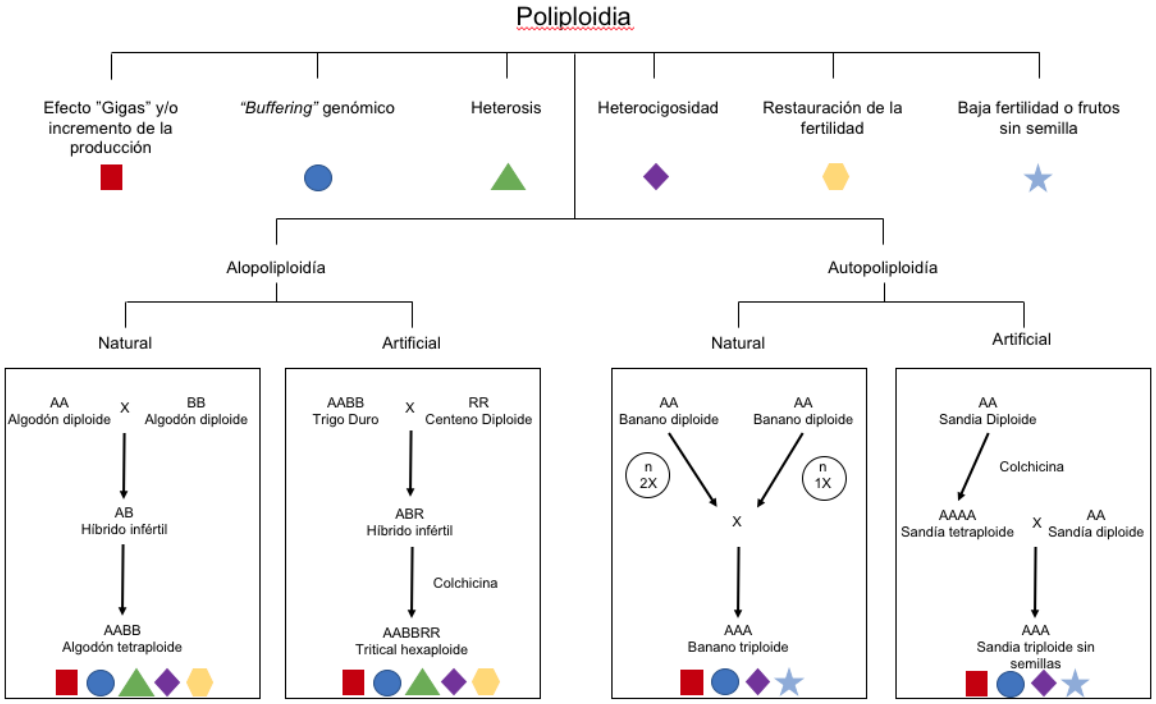


Figura 1. La poliploidía y su efecto en el genoma de las plantas.

Modificada de: (Sattler, Carvalho, & Clarindo, 2015a).

2.1.1.2. Mapeo

La secuenciación del genoma humano, el genoma de *Caenorhabditis elegans* y el genoma de *Arabidopsis*, mostraron la utilidad de la secuenciación BAC por BAC, la cual está basada en la identificación de un camino de recorrido mínimo de clones ó MTP que abarca una región genómica determinada. Por esta razón un mapa físico integral es esencial para este enfoque de la secuenciación del genoma. Además, una correlación establecida entre los mapas físicos y genéticos también es esencial para realizar la clonación de genes y asociar genes candidatos con rasgos biológicos o agronómicos importantes (M. Chen, Presting, & Plant, 2002).

2.1.1.2.1. Mapa Genético

El mapeo genético es el proceso de determinar el orden y la distancia relativa entre los marcadores genéticos (secuencias específicas o elementos heredables que generan un fenotipo) en un cromosoma en función de su patrón de herencia (Nature, 2018).

El mapeo genético en los cultivos generalmente se lleva a cabo en poblaciones segregantes de mapeo, como las poblaciones F2, *Recombinant Inbred Lines* (RIL)[2] y líneas

[2] Líneas recombinantes puras.

retrocruzadas[3]. Los pasos posteriores a este consisten en realizar el mapeo fino y la clonación génica, usando poblaciones avanzadas derivadas de retrocruzamiento. Como un ejemplo de esto, en arroz, una población de mapeo derivada de un cruce entre *Oryza sativa* ssp. indica var. Kasalath y *O. sativa* ssp. japonica var. Nipponbare permitió la identificación y clonación de decenas de QTLs subyacentes a una amplia gama de rasgos (Harushima *et al*., 1998). Aunque esta estrategia se ha utilizado exitosamente en estudios de genómica funcional en cultivares, existen dos limitaciones principales para el mapeo de QTLs en poblaciones recombinantes. Primero, solo hay algunos eventos de recombinación en la población de mapeo; por ejemplo, típicamente ocurren una o dos recombinaciones por cromosoma en poblaciones de segregación, lo que llevaría a una mala resolución de mapeo a menos que se usen poblaciones muy grandes. En segundo lugar, debido a que en la divergencia de secuencias entre los padres seleccionados en una población segregante se representa solo una pequeña fracción de toda variación genética dentro de una especie, ya que en una sola población de mapeo segregante, solo se pueden detectar QTLs en los que difieren los dos padres (X. Huang & Han, 2014).

Para superar estas desventajas, se han construido y utilizado algunos nuevos tipos de poblaciones. Por ejemplo, en maíz, el *Nested Association Map* (NAM) se desarrolló para permitir alta potencia y alta resolución a través del análisis de asociación de cruces articulados (McMullen *et al*., 2009). La población NAM fue creada al cruzar 25 líneas endogámicas de maíz diversas hacia la línea de referencia B73, donde la población NAM en total incluye ~5,000 RIL. La población NAM se ha utilizado en el mapeo genético a gran escala para varios rasgos importantes en maíz (Tian *et al*., 2011). En el caso de la planta modelo Arabidopsis, se creó la población de generación avanzada multiparental (MAGIC, por sus siglas en inglés), que incluyó cientos de RIL descendientes de un grupo heterogéneo de 19 accesiones de Arabidopsis. Para la planta modelo Arabidopsis, se produjo un conjunto de seis poblaciones de RIL denominadas población multiparental de Arabidopsis RIL (AMPRIL). El análisis de QTL en la población de AMPRIL mostró que este recurso genético fue capaz de detectar QTLs explicando el 2% o más de la variación de un rasgo (X. Huang & Han, 2014).

Como una alternativa al análisis de ligamiento convencional, se encuentra el análisis de segregación en masa acoplado a la secuenciación de agrupación de muestras múltiples que se pueden usar en el mapeo genético de rasgos cualitativos simples, o en el mapeo de mutantes. Se han reportado varios métodos para esta aplicación, incluyendo SHOREmap, mapeo de próxima generación, MutMap y MutMap-Gap. En Arabidopsis, se detectaron los sitios de mutación causal cruzando la referencia mutante Columbia (Col-0) con la accesión de *Landsberg erecta* (Ler-1), secuenciandose un grupo de 500 plantas mutantes F2. Para evitar posibles interferencias de diferentes orígenes genéticos, se recomendó el uso de poblaciones derivadas de retrocruzar la línea mutante al padre no mutado, para el mapeo por secuenciación (X. Huang & Han, 2014).

- *Marcadores Moleculares*

Un marcador molecular es un sitio de heterocigosidad para algún tipo de variación del Ácido Desoxirribonucleico (ADN) no asociada con ninguna variación fenotípica medible. El marcador, en este caso tomado como locus, cuando es heterocigótico, puede usarse en el análisis de mapeo tal como se puede usar un par alélico heterocigótico

[3] Un cruce entre el híbrido F1 y uno de sus padres se conoce como retrocruzamiento.

convencional. Debido a que los marcadores moleculares se pueden detectar fácilmente y son numerosos en un genoma, cuando se mapean mediante un análisis de ligamiento, llenan los vacíos entre genes de fenotipo conocido. En el mapeo, el significado biológico del marcador de ADN no es importante en sí mismo; es simplemente un punto de referencia útil para encontrar el orden de la información contenida en los cromosomas (Griffiths *et al*., 2000).

La genotipificación, es la aplicación de los marcadores moleculares, la cual implica la identificación de polimorfismos alélicos en el ADN. Estos polimorfismos son descritos usando una combinación de marcadores moleculares (basados en PCR y/o basados en enzimas de restricción) y el análisis de la secuencia del ADN. Los marcadores moleculares basados en PCR incluyen RAPDs (DNA polimórfico amplificado aleatoriamente) y SSR (*Simple Sequence Repeats*). Los RFLP (polimorfismos de longitud de fragmentos de restricción) es un ejemplo de marcador molecular basado en enzimas de restricción, y marcadores tales como CAPS (secuencias polimórficas amplificadas segmentadas) y AFLP (polimorfismos de longitud de fragmentos amplificados) que usan enzimas de restricción y amplificación de PCR para producir marcadores moleculares nuevos. La genotipificación basada en secuencias de ADN puede tomar la vía de secuenciación directa de genes y regiones genómicas para identificar diferencias, con regiones genómicas asociadas a filogenia, las cuales incluyen regiones ITS (espaciador del transcrito interno), el rADN (ADN ribosomal) y los cloroplastos. Los SNP (polimorfismos de nucleótido único) también pueden identificarse en la secuencia de ADN y usarse como alelos marcadores para la genotipificación. Todos los métodos que dependen de la amplificación de los alelos polimórficos a partir de la secuencia genómica pueden verse afectados por la poliploidía de la especie en estudio, lo cual tendría un efecto comparable con el grado de similitud en la secuencia entre las regiones genómicas duplicadas (Mason, 2014).

- *Quantitative Trait Loci (QTL)*

Un rasgo cuantitativo es aquel que tiene una variación fenotípica medible debido a influencias genéticas y / o ambientales. Esta variación puede consistir en valores discretos, tales como el número de tumores separados en el intestino de un ratón propenso al cáncer, o puede ser continua, tales como las mediciones de la altura, el peso y la presión arterial. (Paran & Zamir, 2003).

Un QTL es un locus genético, y sus alelos que afectan esta variación. En general, los rasgos cuantitativos son multifactoriales y están influenciados por varios genes polimórficos y el ambiente, por lo que más de un QTL puede influir en un rasgo o un fenotipo. Es importante recordar que la variación fenotípica también puede ser causada por factores ambientales que son independientes del genotipo o por medio de interacciones G × A (Genotipo × Ambiente). Genes polimórficos estrechamente ligados pueden ser responsables de la variación cuantitativa de un rasgo, los cuales son difíciles de separar por eventos de recombinación y por lo tanto podrían ser detectados como un QTL. Sin embargo, si son caracterizados como QTLs distintos, separados por medios genéticos o funcionales, cada uno debe ser considerado como un QTL separado (Mackay, 2004).

La variación en los alelos QTL en las plantas ha sido identificada tanto en regiones codificantes como reguladoras de genes individuales, de forma similar a la variación identificada en numerosos genes que controlan los rasgos cualitativos. Un QTL, por lo tanto, puede considerarse como una etapa intermedia en un análisis genético, entre un locus estadísticamente definido y un gen mendeliano. Las potentes herramientas genéticas

desarrolladas en las plantas han demostrado que esta transición de un QTL a un gen mendeliano es factible no solo para QTLs mayores, sino también para los más pequeños. La capacidad de asociar QTLs con secuencias ha conducido a la revelación de nuevas funciones para genes conocidos y al descubrimiento de fenotipos para genes de función desconocida. Los QTLs de plantas que se mapean a alta resolución tienen como objetivo avanzar hacia asociaciones de fenotipos complejos con sus factores subyacentes, ayudando en el estudio de la redes biológicas que interactúan para crear un fenotipo (Paran & Zamir, 2003).

2.1.1.2.2. Mapa físico

Los mapas genéticos proporcionan marcadores a lo largo de los cromosomas. Sin embargo, a menudo hay grandes distancias entre los marcadores para proporcionar la ubicación exacta de los genes. La relación de kilobases por centiMorgan (kb / cM) es grande incluso en plantas modelo como Arabidopsis en la cual la relación entre kilobases y centimorgans es de 120- 250 kb/cM, y en plantas como maíz es aún mayor con 500-1500 kb/cM. En consecuencia, un intervalo de 1 cM puede variar entre ~ 30-100 genes o más. Los mapas físicos se utilizan para conocer la distancia que representa el fragmento completo de ADN entre marcadores moleculares vecinos (Campos-De Quiroz 2002). Los mapas físicos de primera generación estaban basados en cromosomas artificiales de levadura (YAC por sus siglas en ingles), los cuales presentaban problemas de quimerismo y estabilidad, llevando a la introducción y desarrollo de cromosomas artificiales bacterianos (BAC) (Shizuya *et al*., 1992, Peterson *et al*. 2000). Los primeros genomas eucariotas se secuenciaron utilizando un enfoque de mapeo físico (Peterson 2014), utilizando los clones BAC para el ensamblaje de genomas con el fin de obtener la secuencia completa. A partir de estos, se desarrollaron métodos de investigación para la determinación del tamaño del genoma, el mapeo físico y el mapeo genético que permitió la caracterización y la comparación de los genomas, además de emplearlos en la validación, corrección y explotación de los datos de la secuenciación de ADN antes del advenimiento de las tecnologías NGS (*Next Generation Sequencing*) (Peterson 2014). La llegada de las nuevas técnicas de secuenciación ha permitido secuenciar, ensamblar y analizar los genomas de muchas especies de plantas, permitiendo la transformación del análisis molecular en las plantas de un solo gen a todo el genoma (Flagel & Blackman 2012). La información recopilada a través del análisis de los datos de la secuencia del genoma completo puede aplicarse para determinar la función y la regulación de los genes, para el análisis de las relaciones evolutivas entre organismos, soportando el estudio sistemático de la organización del genoma y de la biología de las plantas (Soneji *et al*., 2010).

- *FPC (Finger-Printed Contigs)*

El FPC es un programa interactivo para crear *contigs* a partir de clones con perfiles de identificación del ADN o "Huella Dactilar" (*finger-printed*). El programa proporciona características avanzadas para el montaje, visualización y edición de *contigs*. Además, el programa FPC puede confirmar superposiciones y buscar *contigs*, mediante el uso de marcadores (Soderlund, Longden, & Bioinformatics, 1997).

La huella dactilar es una técnica de comparación de clones basada en la coincidencia de conjuntos de fragmentos característicos entre clones. Este método toma huellas dactilares al usar conjuntos de fragmentos generados para los clones proporcionandoles dirección con dos enzimas de restricción y marcando (radiactiva o fluorescentemente) el extremo de uno

de los sitios de la enzima de restricción. Los fragmentos son separados por electroforesis en geles de poliacrilamida determinando la distancia de migración de los fragmentos marcados, los cuales son llamados bandas. Las bandas representan el subconjunto del ADN contenido en el clon. Por ejemplo, en el proyecto *C. elegans*, los clones de cósmidos se digierieron usando una reacción *Hin*dIH / *Sau*3A donde solo se marcaron los sitios *Hin*dIII, produciendo un promedio de 23 bandas por clon con valores de distancia que generalmente oscilan entre 200 y 3500. Usando este método, dos bandas se consideran iguales si sus valores están dentro de un rango constante. Para que un *contig* sea adecuadamente asignado a una localización, debe tener en una profundidad de cuatro o cinco clones que son canónicos a exactos o aproximados indicando que tienen buenas bandas y son confiables (Soderlund *et al*., 1997).

2.1.1.2.3. Aplicaciones del Mapeo

Existen varios ejemplos de estudios donde se ha realizado el mapeo para obtener información aplicando el mapeo genético y físico. Dentro de estos, se reporta el caso del genoma del arroz, el cual fue utilizado para dilucidar un QTL asociado a la producción, como una citoquinina oxidasa. Así mismo, se identificó un factor de transcripción que controla la expresión del gen, el cual se reconoció como un factor de la tolerancia a la sequía y la sal, *Drought and Salt Tolerance* (DST). Del mismo modo, la secuencia del genoma del maíz permitió desarrollar mapas de haplotipos y encontrar QTLs asociados a la biomasa y la bioenergía (Bolger *et al*., 2014).

Otro caso exitoso es el del genoma del tomate donde se identificó una esterasa responsable de las diferencias en el contenido de éster volátil en diferentes especies. También se identificó un gen que subyace en el locus de maduración uniforme en el tomate, que resultó ser un factor de transcripción tipo Golden 2 que determina la distribución de la clorofila en frutos no maduros. El genoma junto con el borrador del genoma de su pariente silvestre *Solanum pennellii* dejó entrever la evolución de la biosíntesis del terpeno. (Bolger *et al*., 2014).

Los rasgos con valor agronómico incluyen el rendimiento del cultivo (semilla, grano, etc.), el control de la floración y el tiempo de maduración. Estos últimos representan rasgos agronómicos para la adaptación a diferentes regímenes de fotoperíodo y latitud geográfica. Por ejemplo, la secuencia del genoma de la papa ha ayudado en la identificación de un factor de transcripción que regula la madurez de la planta y el ciclo de vida. El genoma de la remolacha azucarera fue usado para determinar la biología del control del tiempo de floración, conocimiento que puede favorecer la creación de una "remolacha invernal" evitando la inducción de las estructuras de resistencia durante el invierno, lo cual es útil para prolongar la temporada de crecimiento con el fin de obtener mayores rendimientos (Bolger *et al*., 2014).

2.1.2. Secuenciación

En el 2002 se publicó la secuencia del genoma del arroz, la primera planta cultivada en ser secuenciada (Michael & Jackson, 2013). Desde entonces, la cantidad de genomas de cultivos secuenciados cada año ha aumentado constantemente (225 para 2017). El ritmo al que las secuencias de genomas son generadas se debe en gran parte a la mejora en las tecnologías de secuenciación tanto en términos de costo como de velocidad. Las nuevas

tecnologías permiten la secuenciación de genomas de plantas cultivadas a un costo razonable. Aunque muchos de los genomas publicados se consideran incompletos, han demostrado ser una herramienta valiosa para comprender rasgos importantes de los cultivos, como la maduración de la fruta, los rasgos del grano y la adaptación del tiempo de floración (Bolger *et al*., 2014).

2.1.2.1. Clases de secuenciación

2.1.2.1.1. Targeted Sequencing

Existen diversos métodos para la secuenciación de un fragmento específico de ADN además del utilizado en este estudio. Un método, es la captura de microarreglos, el cual utiliza la hibridación con matrices que contienen oligonucleótidos sintéticos que coinciden con la secuencia diana para capturar fragmentos de ADN genómico con un adaptador cortado aleatoriamente; se ha aplicado a la secuenciación de 4'200,000 exones. Esta técnica funciona para fragmentos de ADN genómico de 500 bases de largo, lo que limita el enriquecimiento y la eficacia de secuenciación para objetivos muy cortos, como los exones humanos que codifican proteínas que tienen un tamaño medio de 120 bp (Gnirke *et al*., 2009).

Otro método reportado en la literatura usa la amplificación múltiplex. Esta técnica utiliza oligonucleótidos que se sintetizan en un microarreglo. Sin embargo, ha mostrado una representación muy desigual de los objetivos de secuenciación, una reproducibilidad muy pobre entre las réplicas técnicas y una recuperación desigual de los alelos. Un estudio más reciente no basado en secuenciación que utiliza un enfoque similar sugiere que la uniformidad, la reproducibilidad y la eficacia de la amplificación múltiple pueden mejorarse (Gnirke *et al*., 2009).

La secuenciación del genoma de la planta modelo Arabidopsis en el 2000 fue un hito importante no sólo para la investigación en plantas sino también para la secuenciación del genoma. Este hito fortaleció y aceleró la investigación de plantas. La estrategia de secuenciación escogida para esta especie se basó en la superposición de clones de cromosomas artificiales bacterianos (BAC), los cuales representaron una ruta para cubrir cada cromosoma ó MTP. Las secuencias BAC se ensamblaron individualmente y se organizaron de acuerdo con el mapa físico, creando una secuencia de genoma de muy alta calidad. Uno de los inconvenientes que presenta el uso de esta técnica es el costo, esfuerzo y el tiempo, los cuales son muy altos limitando su aplicación a unos pocos genomas. A pesar de estos inconvenientes, tan solo dos años después se obtuvo la secuencia del primer genoma de una planta cultivada, el arroz, el cual es aproximadamente tres veces más grande al genoma de Arabidopsis, utilizando la secuenciación BAC por BAC (Bolger *et al*., 2014).

2.1.2.1.2. Secuenciación de Genoma Completo

El álamo (*Populus trichocarpa*) fue el siguiente genoma de una planta cultivada secuenciada (2006), pero esta empleó la estrategia de secuenciación del genoma completo (WGS) (Michael & Jackson, 2013). En esta estrategia, el genoma se divide aleatoriamente en piezas más pequeñas que luego se secuencian y posteriormente se ensamblan. Dentro de las ventajas que tiene esta estrategia comparada con la secuenciación BAC por BAC, está la reducción del tiempo y el costo de preparación. Sin embargo, los datos son más difíciles de ensamblar y por lo general, dan como resultado una secuencia del genoma más fragmentada. El análisis de otros genomas de plantas cultivadas empleó ambas estrategias, equilibrando

las diversas desventajas y ventajas. En 2007, la secuencia del genoma de la uva fue publicada por dos grupos diferentes, ambos basados en la estrategia WGS. Un equipo se centró en una línea endogámica, mientras el otro analizó un genotipo heterocigótico utilizado para la producción de vino. Estos datos contribuyeron a la evidencia que sugiere estos cultivares pueden ser genéticamente idénticos o no, o pueden mostrar una relación clonal con mutaciones individuales altamente relevantes a nivel fenotípico. Dada la importancia comercial de la uva, la secuencia del genoma ha sido usada para comprender el desarrollo de la uva, identificar factores genéticos que afectan la calidad del vino y producir genotipos que soporten la enorme cantidad de fungicida necesaria para permitir la cosecha, mientras se mantiene la calidad individual de los cultivares de uva (Bolger *et al*., 2014).

La WGS avanzó con la aparición de las tecnologías llamadas secuenciación de nueva generación (NGS) mejorando la razón Salida/Costo. La visión de estas tecnologías es la de secuenciar fragmentos múltiples de ADN en paralelo para producir una mayor cantidad de lecturas, esto a pesar de que son más cortas y con una calidad más baja. En cultivos como cebada, trigo y centeno, los genomas están compuestos por secuencias repetitivas que representan del 80 al 90 % del genoma y están distribuidas a lo largo de este, por esta razón, la secuenciación NGS da como resultado un ensamblaje muy fragmentado, ya que el pequeño tamaño de las lecturas hace que los ensamblajes se vean supeditados a regiones de bajo número de copias dejando las regiones repetitivas e intergénicas sin ensamblar (Bolger *et al*., 2014).

Las plataformas de tercera generación (TGS por sus siglas en inglés) proveen lecturas más grandes las cuales pueden contener y así sobrepasar la limitación de las regiones repetitivas, facilitando el ensamblaje de genomas como el trigo. Además de esta ventaja, las TGS generalmente no requieren de amplificación de ADN evitando el sesgo de la amplificación clonal y producen lecturas en el orden de miles de pares de bases. A pesar de todas estas cualidades, tienen una tasa de error más alta si es comparada con los datos producidos por Illumina. En el presente existen dos tecnologías de tercera generación en producción, PacBio y Oxford Nanopore. (Bolger *et al*., 2014).

2.1.2.1.3. Secuenciación de Genoma completo Vs BAC por BAC

Para genomas grandes, repetitivos y complejos, como es el caso de los genomas poliploides, el ensamblaje de genoma completo (WGS) entrega ensamblajes altamente fragmentados que requieren un considerable esfuerzo y tiempo para ordenar, orientar y así producir pseudo-moléculas aceptables. Para obtener un buen resultado, deben ser incluidas bibliotecas *Mate Pair* (MP) y/o datos de lectura larga de TGS. Estos enfoques deben incluirse para la producción de genomas ensamblados de alta calidad, el uso de mapas genéticos y físicos, la sintenia a especies estrechamente relacionadas con genomas de referencia, genotipificación por secuenciación y secuenciación poblacional. El enfoque tradicional de secuenciación BAC-por-BAC del genoma generalmente producen ensamblajes de mayor calidad que WGS, sin embargo, la secuenciación de BACs es costosa debido a la creación de las bibliotecas BAC, la generación de las huellas digitales de los clones BAC para establecer el MTP y la secuenciación de grandes cantidades de BACs (Visendi et al., 2016).

Como un ejemplo de genomas grandes y complejos que son difíciles de ensamblar están el trigo y la caña de azúcar. El trigo tiene un genoma hexaploide (2n = 6x = 42) que contiene tres genomas diploides ancestrales (AABBDD), cada uno con 7 cromosomas. El genoma es grande, alrededor de 17 Gb y está formado predominantemente por elementos

repetitivos (Zimin et al., 2017). Las variedades de caña de azúcar tienen genomas más pequeños, alrededor de 10 Gb, pero la mayoría son híbridos de dos especies, *Saccharum spontaneum* y *Saccharum officinarum*, siendo *S. officinarum* un octoploide con 2n = 80 cromosomas. *S. spontaneum* número 2n de cromosomas que van desde 40 hasta 128 mostrando una variación en el nivel de ploidía (5-16x) (Hoang *et al*., 2017).

La complejidad de los genomas de trigo y caña de azúcar hace que la producción de genomas de referencia sea difícil. Si bien los enfoques de BAC-por-BAC ofrecen una reducción en la complejidad, los BAC individuales aún contienen un alto porcentaje de repeticiones (~ 80%) en el trigo. Por lo tanto, los BAC generados para los genomas de este tipo son complejos y repetitivos a pesar de la reducción de su tamaño. Disminuir el costo de la secuenciación de BACs mientras se mantiene o mejora la precisión del ensamblaje tiene el potencial de disminuir significativamente el costo de la secuenciación de estos genomas y permite la aplicación de enfoques BAC-por-BAC para diversas especies. Visendi *et al* en 2016, presentaron un enfoque eficiente y rentable para la secuenciación y ensamblaje de BAC complejos utilizando un conjunto optimizado de BACs, para la generación de datos y el ensamblaje bioinformático, demostrando que el uso de este enfoque mediante el ensamblaje de BAC puede llegar a ser útil y eficiente. Por lo tanto, los problemas con esta estrategia, se relaciona más con el ensamblaje y no dependen de la elección específica de la biblioteca usada para secuenciar (Visendi *et al*., 2016).

2.1.2.2. Tecnologías de Secuenciación

Entre las primeras decisiones al comenzar un proyecto de secuenciación de genoma se encuentra la elección de la plataforma de secuenciación, el tipo y la cantidad de datos para generar. Este última suele estar limitada por el dinero disponible para este fin, y el primero puede depender de la tecnología de secuenciación disponible en el mercado. En los últimos años, ha habido progreso en la producción de lecturas más largas y de alto rendimiento; varias tecnologías que ofrecen esto, como Pacific Biociences (hasta 25 kb), IonTorrent (~ 500 bp), Illumina Moleculo (hasta 10 kb), Oxford Nanopore (tamaño limitado por la tecnología de extracción de ADN) se están consolidando en el mercado. Aunque este desarrollo desdibuja la dicotomía inicial de lecturas cortas (por ejemplo, lecturas de Illumina de 35 pb) frente a lecturas largas (~ 1 kb lecturas de Sanger), la longitud de lectura aún tiene importantes implicaciones bioinformáticas, ya que los algoritmos de ensamblaje optimizados para lecturas largas son fundamentalmente diferentes de los enfoques de lectura corta. Los estudios recientes combinan datos de diferentes longitudes de lectura y de varias plataformas de secuenciación. Esta estrategia tiene un sentido intuitivo, ya que los inconvenientes de cada método pueden contrarrestarse mutuamente (Ekblom & Wolf, 2014).

Para la mayoría de las aplicaciones posteriores, la obtención de *scaffolds* largos es esencial. Con lecturas a partir de la secuenciación Sanger tradicional de clones de BAC individuales, es factible obtener un buen ensamblaje incluso con una profundidad de secuencia bastante limitada. Sin embargo, cuando se usan sólo tecnologías de lectura corta, se necesita una cobertura de lectura total alta (> 100X), para evitar un ensamblaje muy fragmentado y problemas con las aplicaciones posteriores, como la anotación y la llamada de variantes. Para el ensamblaje inicial, se comienza con una gran cantidad de datos de lectura corta de extremos emparejados o *paired end*. Para luego fusionar *contigs* en *scaffolds*, es necesario generar bibliotecas adicionales con tamaños de inserción larga en el rango de 3-40 kb. La cantidad de datos de secuenciación necesarios para cada tipo de biblioteca y tamaño de inserción depende críticamente de varios factores, incluidos el tamaño y el contenido

repetido del genoma, el grado de heterocigosidad y la calidad objetivo del ensamblaje (Ekblom & Wolf, 2014; Lu, Giordano, & Ning, 2016).

2.1.2.2.1. Secuenciación por síntesis CRT Illumina ®.

Es un método dependiente de ADN polimerasa, por terminación reversible cíclica (CRT por sus siglas en inglés). Se define por el uso de un terminador de síntesis de moléculas de ADN, que son similares a los utilizados en la secuenciación de Sanger, en la que el grupo 3'-OH de la ribosa está bloqueado, evitando así la elongación. La plantilla de ADN es iniciada por una secuencia que es complementaria a una región adaptadora, que comenzará la unión de la polimerasa a esta región de ADN de doble cadena (dsADN). Durante cada ciclo, se añade una mezcla de los cuatro desoxinucleótidos marcados individualmente y bloqueados en 3 '(dNTP). Después de la incorporación de un solo dNTP a cada hebra complementaria alargada, se eliminan los dNTPs no unidos y se captura la imagen de la superficie para identificar qué dNTP se incorporó en cada grupo. El fluorocromo y el grupo bloqueante pueden entonces ser eliminados y puede iniciarse un nuevo ciclo (Goodwin, McPherson, & McCombie, 2016).

2.1.2.2.2. Pacbio®

La secuenciación PacBio® captura información de la secuencia durante el proceso de replicación de la molécula de ADN blanco. La plantilla, llamada SMRTbell, es un ADN circular cerrado, monocatenario, que se crea ligando los adaptadores de horquilla a ambos extremos de una molécula diana de ADN bicatenario (ADNdc). Cuando se carga una muestra de SMRTbell en un chip llamado SMRT, ésta se difunde en una unidad de secuenciación llamada guía de onda de modo cero (ZMW), que proporciona el volumen más pequeño disponible para detección de luz. En cada ZMW, una sola polimerasa se inmoviliza en la parte inferior, la cual puede unirse al adaptador de horquilla del SMRTbell y comenzar la replicación. Los cuatro nucleótidos están marcados con fluorescencia, generando espectros de emisión distintos, los cuales se añaden a la celda SMRT. Los procesos de replicación en todas las ZMW de una celda SMRT se registran mediante un video de los pulsos de luz, y los pulsos correspondientes a cada ZMW se pueden interpretar como una secuencia de bases (llamada lectura larga continua, CLR). Debido a que SMRTbell forma un círculo cerrado, después de que la polimerasa replica una cadena del ADNdc, puede continuar incorporando bases del adaptador y luego de la otra cadena. La polimerasa puede secuenciar varias veces (llamadas "pases") un único CLR. En este escenario, el CLR se puede dividir en varias lecturas (llamadas *subreads* o *reads of insert*) reconociendo y eliminando las secuencias del adaptador. La secuencia de consenso de múltiples *subreads* en una única ZMW produce una secuencia de consenso circular (CCS) la cual tienen un porcentaje de error cercano a la secuenciación de fragmento corto. Debido a que la secuenciación de PacBio® tiene lugar en tiempo real, la variación cinética interpretada a partir de la película de pulso de luz se puede analizar para detectar modificaciones de las bases, como la metilación (Rhoads & Au, 2015).

2.1.2.3. Análisis bioinformático para el ensamblaje y anotación de genomas en plantas poliploides

Existen muchos desafíos para el ensamblaje de un genoma poliploide. El enfoque algorítmico común para el ensamblaje del genoma se centra en encontrar rutas lineales en el gráfico de ensamblaje construido a partir de superposiciones entre las lecturas. Cuando

diferentes haplotipos se 'mezclan' y se secuencian juntos, regiones similares entre los haplotipos a menudo se colapsan en trayectorias lineales mientras que los caminos se dividen en regiones diferentes, formando 'burbujas' en el gráfico de ensamblaje fragmentándolo (Figura 2). El gráfico de ensamblaje de un genoma poliploide típico presenta muchas de estas burbujas, a menudo anidadas una dentro de otra. El nivel de variación de secuencia entre los haplotipos (heterocigosidad), así como el patrón de distribución de estas variaciones en el genoma, tiene un gran impacto en la calidad del ensamblaje (Tang, 2017).

En los poliploides, a menudo se hace una elección entre una secuencia consenso que colapsa los haplotipos (monoploide) frente a la reconstrucción por separado de cada haplotipo. La recuperación de los distintos haplotipos se puede realizar mediante un ensamblador de genoma *de novo* compatible con haplotipos junto con la secuencia de lectura larga. Este enfoque es adecuado cuando los haplotipos son divergentes, particularmente con un alto grado de variación estructural. Sin embargo, cuando el ensamblaje *de novo* no puede distinguir los haplotipos, probablemente debido a la falta de lecturas largas, la estrategia alternativa es tener una representación monoploide con los haplotipos colapsados inicialmente, seguidos de la eliminación gradual de variantes para resolver los haplotipos, como se muestra en la Figura 2 (Tang, 2017).

Para resolver los haplotipos de batata, se utilizó un nuevo método de eliminación gradual, denominado *Ranbow*. El método comenzó desde un conjunto inicial de variantes que forman haplotipos 'semilla', y luego extendió los haplotipos cuando fue posible. Para evaluar la precisión, los haplotipos resueltos fueron comparados contra un conjunto separado de lecturas largas. La validación adicional también mostró que estos haplotipos cubren en su mayoría el mismo locus genético donde se encuentran secuencias repetitivas, lo que impide un tipo común de error durante el ensamblaje del genoma (Tang, 2017).

La resolución de haplotipos o *phasing*, presenta limitaciones. La primera es que los sitios adyacentes a las variantes deben estar contenidos dentro de una lectura para que puedan ser identificados (*phased*). Por lo tanto, si el nivel de heterocigosidad es demasiado bajo en comparación con la longitud de lectura, los bloques de fase podrían estar limitados por el tamaño. Una segunda limitación, consiste en que este sólo funciona en variantes pequeñas, como polimorfismos de un solo nucleótido e inserciones y eliminaciones cortas ya que las variantes estructurales de traslocación, o duplicación segmental entre los haplotipos puede ser ignorada ya que los fragmentos que permiten realizar el *phasing son* de tamaño limitado, inclusive con las nuevas tecnologías de secuenciación de fragmento largo, aunque mitigan el efecto de fragmentación del ensamblaje. A pesar de todos estos inconvenientes, el método muestra potencial para su uso en genomas poliploides, especialmente genomas auto-poliploides que son actualmente menos asequibles para ensamblar. La mayoría de los métodos de *phasing* usan árboles filogenéticos, los cuales pueden o no soportar los diferentes haplotipos. Dentro de los procesos biológicos que tienden a homogenizar los haplotipos, y por ende dificulta el proceso de *phasing* se encuentran la recombinación, el intercambio homólogo, la conversión de genes y la introgresión, los cuales son conocidos y reportados en genomas poliploides. Así mismo, los errores del *phasing* de los haplotipos junto a los procesos descritos anteriormente, pueden explicar los resultados extraños obtenidos en algunos árboles genéticos usados para desarrollar el proceso de identificación de los haplotipos (Tang, 2017).

Las técnicas de mapeo del genoma como el Hi-C o el mapeo óptico, así como la secuenciación de lecturas largas, como PacBio® o Oxford Nanopore, podrían mejorar el proceso de *phasing* que permita la reconstrucción de ensamblajes a escala cromosómica. Con

los avances en la tecnología de secuenciación y mapeo, así como en el desarrollo algorítmico, se espera que en el futuro cercano sea más fácil el ensamblaje de genomas poliploides (Tang, 2017).

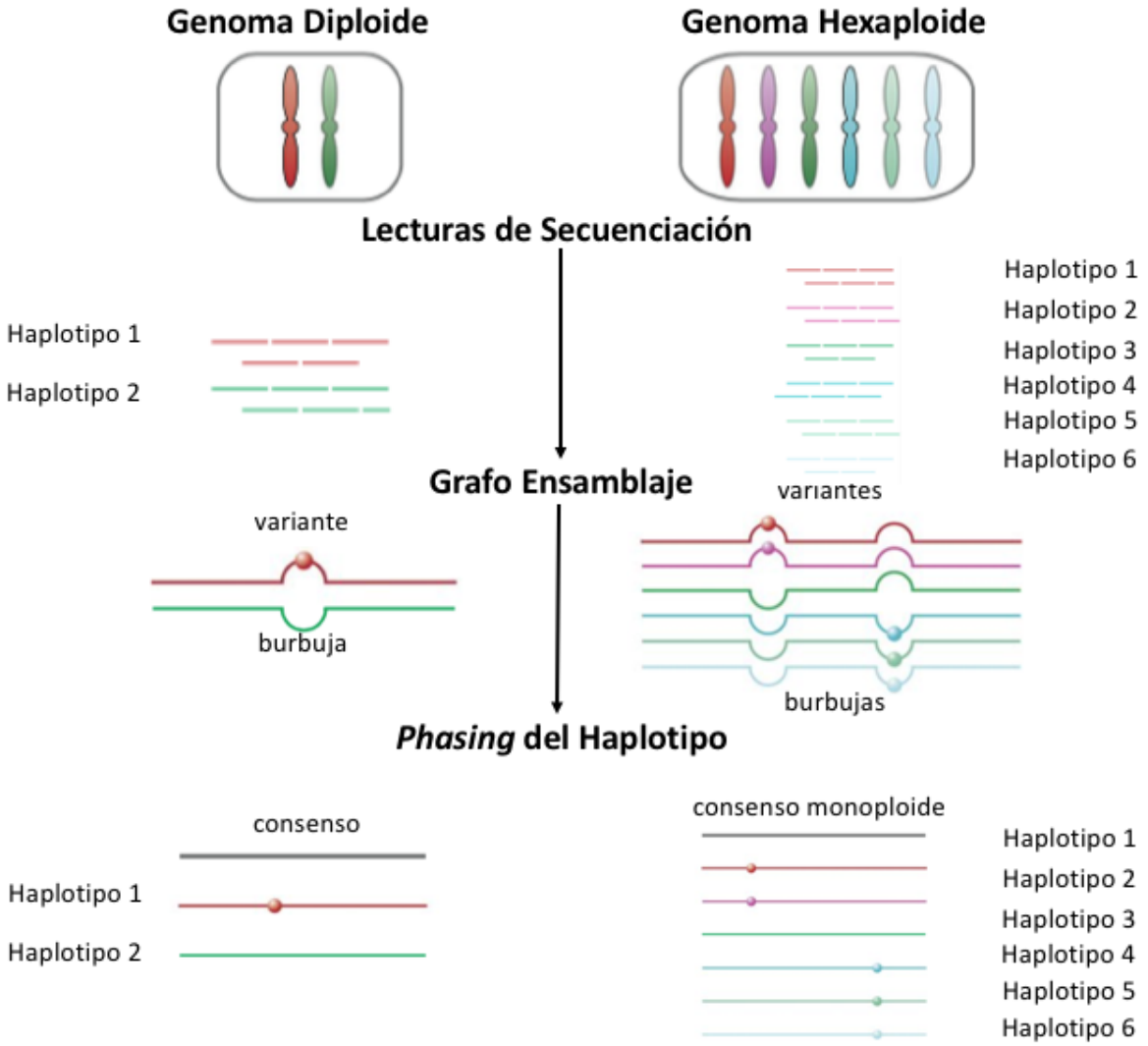


Figura 2. Ensamblaje del genoma y *phasing* de haplotipos. Un diploide con dos haplotipos (izquierda) y un hexaploide con seis haplotipos (derecha). Múltiples haplotipos se secuencian juntos y las lecturas se pueden derivar de cada uno de los haplotipos. El gráfico de ensamblaje luego se construye sobre la base de las superposiciones entre las lecturas, donde los sitios heterocigotos entre los haplotipos forman 'burbujas' en el gráfico. Los haplotipos a veces se colapsan en un consenso como referencia y es necesario un paso de fase de separado para recuperar los haplotipos subyacentes dentro del genoma Modificado de: (Tang, 2017).

Las tecnologías TGS ayudan en el mejoramiento de la calidad de las secuencias de referencia de alta calidad y se alejan de los ensamblajes de lectura corta de bajo costo. Los ensambles de lectura larga *per se*, superan a los ensamblajes de lectura corta y en combinación con el mapeo óptico u otros datos de largo alcance surge como estrategia para

ensamblar genomas con una alta integridad y calidad. Esto podría marcar una tendencia, donde la dependencia por un genoma de referencia será mucha más baja debido a la disponibilidad de muchos ensamblajes de buena calidad.

Mientras que los estudios se concentran en la generación de referencias específicas, cada vez es más frecuente encontrar algoritmos que pueden manejar múltiples secuencias de referencia para soportar alineamientos simultáneos contra múltiples referencias o para facilitar la identificación de polimorfismos dentro de los gráficos de ensamblaje a partir de datos de secuencia de múltiples individuos (Tang, 2017).

A pesar de todas las ventajas que pueden suministrar las tecnologías de secuenciación de tercera generación, los proyectos que implican secuenciación a escala poblacional todavía usan tecnologías de lectura corta Illumina, por lo tanto, estos estudios deben lidiar con todas las deficiencias de la resecuenciación, como el sesgo a la referencia, aún generando secuencias de referencia específicas para cada trabajo. El advenimiento de técnicas como la plataforma 10X *Genomics* tiene el potencial de revolucionar los proyectos de secuenciación a escala poblacional, disminuyendo el sesgo que suele representar el ensamblaje de datos NGS, haciendo que los datos de lectura corta sean agrupados mediante un *barcode* permitiendo hacer la deconvolución de lecturas cortas a fragmentos largos de > 100kb de los cuáles se realizan las librerías, obteniendo ensamblajes *de novo* de alta calidad para cada individuo de una población con sólo un pequeño aumento en los costos. En la actualidad, existen pocos métodos para analizar estos ensamblajes de ensamblajes, sin embargo, se está generando una tendencia cada vez más creciente hacia el análisis de genomas completos de grandes paneles de plantas, incluidos híbridos y poliploides. Esto marcará el fin de la era de la resecuenciación y permitirá nuevas comparaciones de genomas completos que revelarán y utilizarán un amplio rango de diferencias entre genomas independientemente de su tamaño o complejidad (Tang, 2017).

2.1.2.3.1. Secuenciación asistida por sub-genomas de cultivos de poliploidía compleja

Las especies de plantas poliploides como el arroz, el trigo y el maíz, son responsables del 42% del suministro de la alimentación humana mientras que otros cultivos como el algodón, la soja y la canola desempeñan un papel importante en la nutrición y el vestuario. Los primeros genomas de cultivos, arroz, maíz y sorgo se secuenciaron utilizando la tecnología de secuenciación Sanger con sus enfoques WGS y BAC por BAC. La mayoría de los cultivos son poliploides, lo que probablemente contribuye a su contenido nutricional mejorado y altos rendimientos. A la poliploidía también se le confiere propiedades, como la acumulación de aceite de la semilla en la canola, fibras hilables en el algodón y la composición nutricional del grano en el trigo. La presencia de múltiples subgenomas vuelve más complejo el ensamblaje del genoma de los poliploides, pero la estrategia para la secuenciación de los probables subgenomas ha acelerado la comprensión de estos genomas complejos (Renny-Byfield & Wendel, 2014).

Brassica napus (canola o colza), la tercera fuente más importante de aceite vegetal es un alotetraploide de *B. rapa* (nabos y repollo nana) y *B. oleracea* (repollo, brócoli, coliflor, col rizada y otras verduras crucíferas), cuyo fenómeno de poliploidización ocurrió alrededor de 7500-12,500 años atrás. La mayoría de los 20,000 *scaffolds* basados en Illumina y 454 en *B. napus* se asignaron al subgenoma A o C usando lecturas 454 de cada progenitor parental, lo que permitió comparaciones sin precedentes de regiones homeólogas en una especie poliploide. Como resultados de estas comparaciones se observó la pérdida de alrededor de

100 genes de cada subgenoma, cambios sutiles en la regulación epigenética, el intercambio homeólogo y la divergencia de la expresión génica. Se identificó que el proceso de poliploidización en la canola, presenta repetidas duplicaciones del genoma entero, generando una duplicación 72X del genoma de la planta ancestral, observándose expansiones únicas en los genes de biosíntesis del aceite y la pérdida de genes de glucosinolato (El glucosilonato progotrina, puede afectar a individuos con enfermedades asociadas a la tiroides) (Borrill, Adamski, & Uauy, 2015).

El trigo (*Triticum aestivum*) es un alimento básico para el 30% de la población mundial, sin embargo, los recursos genómicos eran limitados debido al gran tamaño de su genoma (17 Gb), hexaploide (2n = 6x = 42), y rico en repeticiones (87%). Para superar estos problemas con la complejidad del genoma, se utilizó la citometría de flujo con líneas de aneuploidía[4] para aislar y secuenciar cada brazo del cromosoma por separado, representando cada brazo 1,3-3,3% del genoma. El ensamblaje final está muy fragmentado y abarca 10,2 Gb con 124.201 genes distribuidos de forma desigual en los tres subgéneros (Borrill *et al.*, 2015). Otro ensamblaje publicado en 2017 (Triticum 3.0) realizado mediante un ensamblaje híbrido, contiene 15 Gb en 279,529 *contigs*, con un N50 de 232 613 pb y el *contig* más largo es de 4 Mb (Zimin *et al*., 2017). Los subgéneros de trigo presentaron una pérdida limitada de genes o reordenamientos, en contraste con la combinación dinámica y la pérdida en los eventos de poliploidía mucho más recientes en *B. napus*. Esto sugiere plasticidad en los eventos posteriores a la duplicación del genoma completo (WGD) (Borrill *et al*., 2015).

En el caso del algodón tetraploide (AADD, *Gossypium hirsutum*) tiene una mayor producción de fibra y calidad que el algodón diploide (*G. barbadense*), una propiedad emergente con QTL mayores en el subgenoma D de *G. raimondii*, que no tiene fibras hilables. El algodón diploide (*G. raimondii*) presenta un incremento de ploidía de 5-6 veces después de separarse del linaje de cacao, comparado este aumento sólo con las Brassicaceae. El algodón tetraploide tiene numerosos intercambios de ADN no recíprocos entre los subgéneros A y D, y cambios coordinados en la expresión génica, incluidos genes mitocondriales nucleares implicados en el transporte de electrones, que probablemente contribuyen a la producción de fibra (F. Li *et al*., 2015).

El café, es el principal cultivo de bebidas del mundo donde con un consumo de más de 2,25 mil millones de tazas de café por día. La mayoría del café proviene de *Coffea arabica*, un alotetraploide altamente heterocigótico. El genoma de referencia del café se generó utilizando una doble asociación haploide de *C. canephora*, uno de los progenitores diploides de *C. arabica* y la fuente del café 'robusta'. El café tiene duplicaciones en tándem de N-metiltransferasas (NMT) que contribuyen a la producción de cafeína, así como de otros genes que ayudan a la composición química de los granos del café como la oletato desaturasa *FAD*2. Las comparaciones de NMT del té y el cacao sugieren que la biosíntesis de cafeína tiene orígenes polifiléticos fruto de la adaptación mediante la selección positiva, convirtiéndose en un fenómeno de evolución por convergencia de rutas metabólicas secundarias codificadas por genes en tándem (Denoeud *et al*., 2014).

2.1.2.3.2. Mejoramiento asistido por genómica

Existen dos tipos principales de reproducción asistida por genómica: (1) MAS (Selección asistida por Marcadores) y (2) GS (Selección Genómica). La MAS, incluye el retrocruzamiento asistido utilizando marcadores moleculares que se mapean dentro de genes

[4] Individuo cuyo número de cromosomas difiere del individuo silvestre.

específicos o QTLs que se sabe están asociados con rasgos o fenotipos deseados para seleccionar individuos que porten alelos favorables para los rasgos de interés y / o descartar aquellos que no lo hacen. La GS, por otro lado, usa todos los datos de marcadores disponibles para una población como predictores de valor de mejoramiento. Específicamente, la GS integra datos de marcadores de una población de entrenamiento con datos de pedigrí fenotípicos recopilados en la misma población para generar un modelo de predicción. El modelo genera valores genómicos de reproducción estimados (GEBV por sus siglas en inglés) para todos los individuos genotipificados dentro de una población. El GEBV sirve como un predictor del desempeño de la planta como padre para el cruce y el avance de generación en un proceso de mejoramiento, en base a la similitud de su perfil genómico con otras plantas que se sabe que tuvieron un buen desempeño en el ambiente objetivo. Antes de que el modelo de predicción se pueda aplicar a una población reproductora, la precisión del modelo generalmente se prueba usando validación cruzada en subconjuntos de la población de entrenamiento. Una vez validado, el modelo se puede aplicar a una población reproductora donde se calculan los GEBV para todas las líneas para las que se dispone de información genotípica, y su rendimiento fenotípico se predice únicamente sobre la base de esa información genotípica (Varshney, Terauchi, & McCouch, 2014).

La ventaja de la reproducción asistida por genómica radica en que los datos genotípicos obtenidos de una semilla o plántula se pueden utilizar para predecir el rendimiento fenotípico de individuos maduros, ahorrando tiempo y dinero donde se hace innecesaria una evaluación fenotípica en diferentes ambientes y en varios ciclos productivos. El uso de reproducción asistida por genómica, tanto en MAS como en GS, permite más ciclos de selección y mayor ganancia genética por unidad de tiempo. (Varshney *et al*., 2014).

2.1.3. Relación genotipo y fenotipo

Una planta está expuesta a las fluctuaciones en las condiciones ambientales, enfrentándose a la heterogeneidad ambiental a la cual debe adaptarse. Uno de los mecanismos de adaptación es el cambio de la expresión fenotípica (plasticidad fenotípica). La plasticidad a menudo significa alterar la expresión génica y la fisiología de la planta en respuesta a las condiciones ambientales (Assmann, 2013; X. Huang & Han, 2014; Juenger, 2013). El efecto diferencial producido por esta plasticidad en plantas ya ha sido reportado y se ha considerado en el desarrollo de cultivares, aunque no deja de ser un desafío. Los avances tecnológicos en genotipificación han dado como resultado un cambio en el enfoque de la investigación en fitomejoramiento, que va desde la generación de marcadores moleculares, hasta la fenotipificación automatizada de alto rendimiento. Con estos avances, se ha hecho más fácil determinar el efecto que tiene el ambiente sobre el fenotipo y, por lo tanto, comprender mejor la plasticidad fenotípica. La plasticidad fenotípica y la etapa de desarrollo específica se describe como Genotipo por Ambiente (G × A) lo cual quiere decir que el fenotipo no sólo depende del genotipo (Kang, 1997; Weigel, 2012).

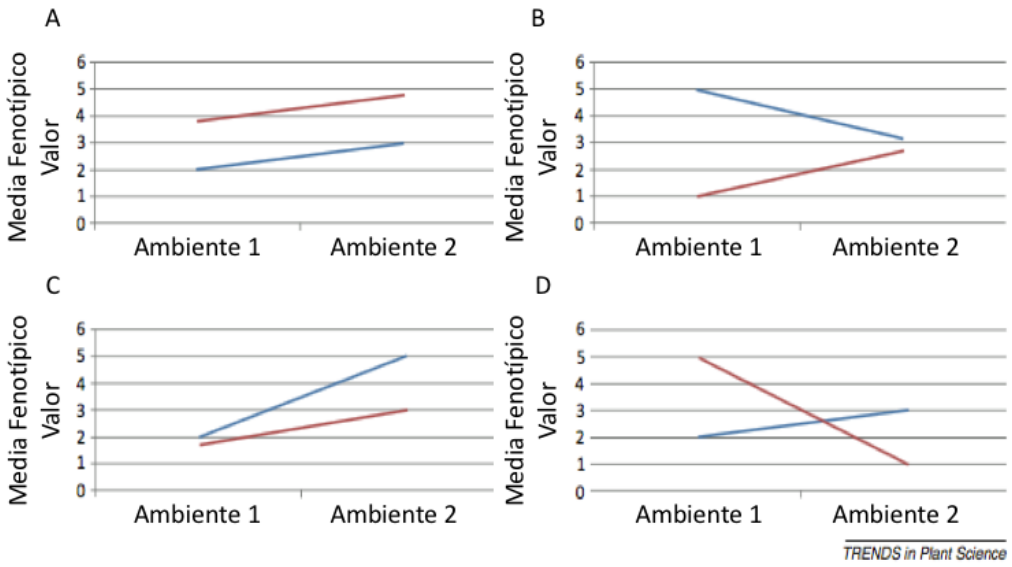


Figura 3. Plasticidad fenotípica e interacción genotipo ambiente (G × A). Cuatro ejemplos de normas de reacción que ilustran (A) la plasticidad fenotípica de los genotipos entre ambientes, y (B-D) la diferencia genética entre los genotipos para la plasticidad que indica G × A. El genotipo 1 es la línea roja y el genotipo 2 la línea azul en cada gráfico. Modificado de: El-Soda *et al*., (El-Soda, Malosetti, Zwaan, Koornneef, & Aarts, 2014).

La respuesta fenotípica al ambiente se puede expresar en las llamadas "normas de reacción"[5] (Figura 3), donde la interacción G × A ocurre cuando la pendiente de las normas de reacción difiere para los genotipos (Figuras 3B-D). Debería hacerse una distinción importante entre el caso más extremo de una interacción cruzada, en la que las normas de reacción se cruzan entre sí (Figura 3D), y una interacción de Efecto de Escala (EE) (Figura 3 B, C), donde no se observa ninguna intersección entre las normas de reacción dentro del rango de los ambientes. Dado que la interacción cruzada implica un cambio en la clasificación de los genotipos en el rango ambiental, se considera que es biológicamente más relevante que la interacción EE (Lacaze, Hayes, & Korol, 2008). En la figura 3-A, los valores del rasgo cambian por ambiente, pero las normas de reacción se ejecutan en paralelo porque la respuesta a las condiciones ambientales es similar para ambos genotipos. Mientras que en la figura 3-B, C hay una interacción de EE que indica la interacción G × A, porque cada genotipo tiene una respuesta diferente a cada uno de los ambientes, pero sin las normas de reacción cruzadas, mientras que en (D) hay un efecto genético más fuerte en el fenotipo en respuesta a los diferentes ambientes, indicando la interacción G × A, provocando que las normas de reacción se crucen La interacción G × A no se limita entre una sola planta o genotipo y sus condiciones ambientales circundantes, debido a que puede extenderse a la interacción entre plantas individuales o entre plantas y sus patógenos en el mismo ambiente, lo que se conoce como interacción Genotipo × Genotipo × Ambiente (G × G × A) (Jorgensen, 2012).

La heredabilidad de un rasgo se calcula en función de los efectos deseados, predeciendo la respuesta genotípica a la selección a través de los ambientes, dejando los efectos de G × A dentro de la varianza fenotípica no explicada, lo que se denomina herencia

[5] Respuesta del fenotipo a los diferentes ambientes.

'perdida' (Chevin, Lande, & Mace, 2010). La falta de heredabilidad es el resultado de todas las fuentes, lo que puede conducir a desviaciones del modelo de efecto aditivo simple, que incluye epistasis, variantes epigenéticas, variantes raras, QTLs pequeños no detectados y la interacción G × A. Evaluar genotipos en múltiples ambientes es esencial para obtener una idea de la extensión de la interacción G × A lo que es de gran interés para los mejoradores ayudándolos a evaluar el progreso de la selección logrado en un ambiente y así determinar que genotipos pueden ser llevados a otros ambientes. (El-Soda *et al*., 2014).

Se han descrito cinco modelos o mecanismos genéticos para explicar la base genética de la interacción G × A:

- Sobredominancia: la plasticidad es una función de la homocigocidad, con una plasticidad fenotípica correlacionada negativamente con el número de loci heterocigotos.

- Pleiotropía o sensibilidad alélica: un gen pleiotrópico afecta dos rasgos plásticos y puede tener diferentes efectos en ambos rasgos en diferentes ambientes.
- Epistasis: la plasticidad producida por loci que puede modificar la expresión de otros genes para activarlos o desactivarlos de una manera específica por ambiente.

- Enlace genético: los alelos que promueven la plasticidad pueden ligarse con los alelos que confieren un fitness bajo o alto.

- Epigénesis: el impacto de la modificación de la cromatina y la metilación del ADN puede depender del medio ambiente y puede conducir a cambios epigenéticos en la función génica y en consecuencia cambios en los fenotipos, sin cambiar la secuencia de ADN.

Dentro de las interacciones fenotipo por ambiente, se ha venido usando con mayor frecuencia la palabra "fenoma", la cual se refiere al fenotipo como un todo (Soulé, 1967). El fenoma de la planta depende del genoma, las condiciones ambientales y sus complejas interacciones dinámicas (G × A) (Houle *et al*., 2010; Chen *et al*., 2014). El manejo del cultivo (M) es un factor adicional que influencia el fenotipo. Por lo tanto, un fenotipo de un cultivar en un sistema agrícola específico se puede considerar como el resultado de las interacciones G × A × M (Porter y Christensen, 2013) que se muestran en la Figura 4. Comprender y explotar esta plasticidad fenotípica en el fitomejoramiento es a la vez desafiante e importante, en particular cuando las plantas cultivadas se ven afectadas con mayor frecuencia por condiciones extremas en condiciones abióticas debido al cambio climático global. Se necesita un compromiso entre aumentar el rendimiento óptimo mientras se mantiene la estabilidad del rendimiento.

Con el establecimiento de métodos no invasivos para la fenotipificación de plantas y su aplicación de alto rendimiento en instalaciones automatizadas a gran escala (Finkel, 2009), la fenómica se aplica en la investigación básica de plantas y la fitogenética, a la espera a que la fenotipificación pueda ser incluída en el mejoramiento genético (Furbank y Tester, 2011). Las técnicas de fenotipificación se desarrollaron para entornos controlados y las investigaciones actualmente buscan su aplicación en campo. (Araus y Cairns, 2014; Langridge y Reynolds, 2015). Aunque la fenómica es análoga a la genómica en cuanto a la

caracterización general del conjunto completo de fenotipos de un genoma, el concepto no es comparable debido a que las variaciones de los fenotipos se presentan de célula a célula y de momento en momento, y por lo tanto nunca se pueden caracterizar por completo. Además, los métodos de fenotipificación de plantas aplicados actualmente aún se encuentran rezagados en comparación con la capacidad para caracterizar los genomas.

A pesar de que la fenómica ha sido denominada como "fisiología de las plantas de alto rendimiento", la mayoría de los casos los enfoques de alto rendimiento todavía se centran en el fenotipo externo y en la medición indirecta de los procesos fisiológicos (Ghanem *et al*., 2015) (Furbank y Tester, 2011) (Figura 4).

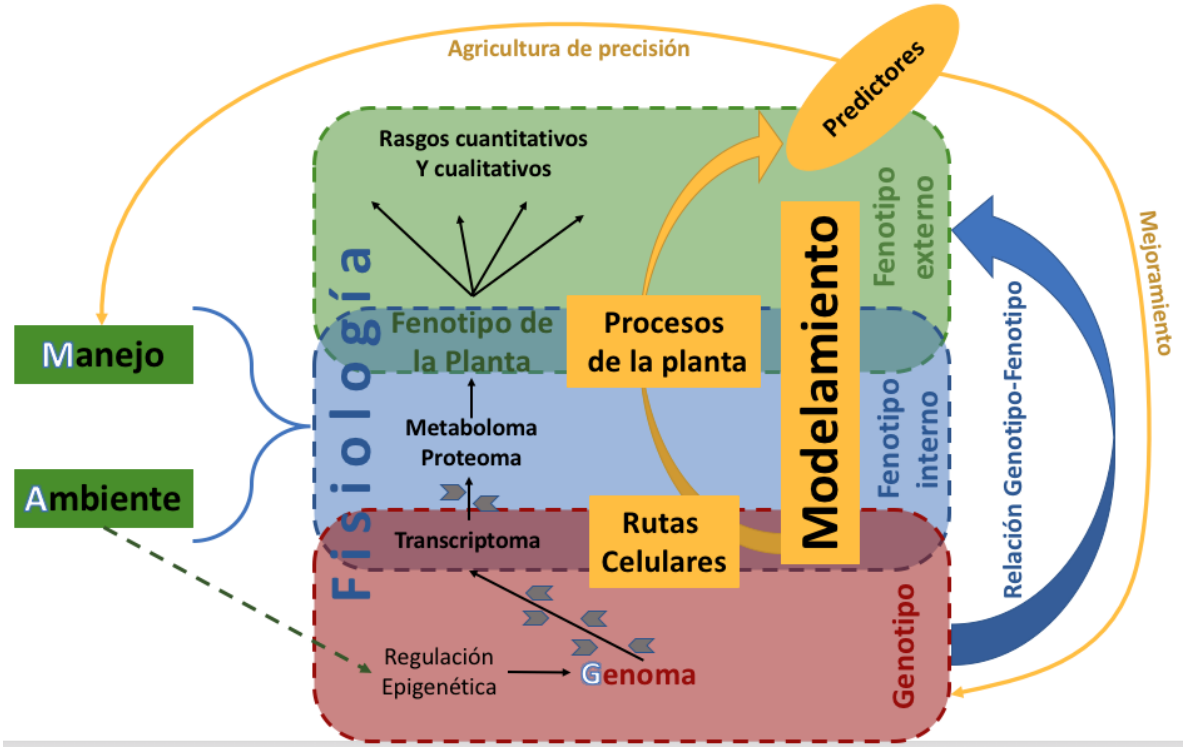


Figura 4. El papel de la fisiología en la interacción (G × A × M). La posición de la fisiología en la integración del genotipo, el ambiente, el manejo y sus interacciones (G × A × M), lo que determina el fenotipo interno como un vínculo entre el genotipo y el fenotipo externo. La fisiología es una interfaz clave entre el genoma y el fenotipo de la planta que comprende múltiples niveles de procesamiento de información a través del transcriptoma, el proteoma y el metaboloma, así como también los eventos regulatorios. La correlación del estado fisiológico como el fenotipo interno con el trasfondo genético y los fenotipos externos, puede alimentar los enfoques de modelado de las vías celulares y los modelos de plantas enteras, lo que permite la identificación y el establecimiento de predictores externos robustos. Tales predictores pueden usarse en aplicaciones prácticas de la agricultura de precisión para cambiar los servicios de gestión o como marcadores para el mejoramiento genético. Modificado de: Großkinsky, Svensgaard, Christensen, & Roitsch, (Großkinsky, Svensgaard, Christensen, & Roitsch, 2015)

A nivel fisiológico, los factores abióticos y bióticos del ambiente multifactorial se integran en las respuestas celulares que determinan en última instancia los rasgos

cuantitativos y cualitativos en el tejido, el órgano y a nivel del organismo completo. Estos procesos metabólicos se pueden monitorear a través de parámetros fisiológicos y bioquímicos que finalmente determinan el fenotipo externo, sin embargo, la resolución está inherentemente limitada por las técnicas usadas actualmente. A pesar de la importancia del enfoque de fenotipos fisiológicos, es importante considerar y definir los aspectos de rendimiento, resolución y dimensionalidad. Por lo tanto, se ha generado la necesidad de implementar a los fenotipos fisiológicos una aproximación holística multidimensional de los fenómenos a través de las distintas escalas de un organismo, que van desde el nivel celular hasta las plantas y los doseles completos de un conjunto de árboles. Este concepto debería ser clave para dilucidar rasgos fenotípicos complejos, en su entendimiento como rasgos fisiológicos específicos y así contribuir a la comprensión de los mecanismos involucrados en la determinación de la plasticidad fenotípica en un complejo entorno multifactorial (Großkinsky *et al*., 2015).

2.2. Café

El café pertenece a la familia Rubiaceae y la tribu Coffeeae y está conformada por más de 124 especies distribuidas en dos géneros, Coffea L. y Psilanthus Hook.f, cada uno de los cuales consta de 2 subgéneros, Coffea y Baracoffea (J.F. Leroy), y Psilanthus y Afrocoffea (Moens), respectivamente. Sin embargo, la producción comercial de café está dominada únicamente por dos especies pertenecientes al género Coffea: *C. arabica* y *C. canephora* (éste último referido generalmente como café robusta) (Charrier *et al*., 2004).

Todas las especies de café son diploides (2n = 2x = 22) y generalmente auto incompatibles, excepto *C. arabica* el cual es un tetraploide autocompatible (2n = 4x = 44) derivado de una hibridación espontánea entre *C. canephora* (progenitor paterno) y *C. eugenioides* (progenitor materno) (Charrier *et al*., 2004, Lashermes *et al*., 1999).

Aunque *C. arabica* es considerado como el café con mayor calidad de taza, aún se continúa investigando en el mejoramiento de la calidad y la producción de las dos especies comerciales (*C. arabica* y *C. canephora*). Con los avances en genómica y las tecnologías de secuenciación, es factible entender el genoma del café y la herencia molecular y así, poder usar estos datos para mejorar la eficiencia de los programas de mejoramiento genético (Tran *et al*., 2016).

2.2.1. Genoma del café

Entre las plantas alopoliploides, *C. arabica* es un alotetraploide reciente (divergencia 100.000 años). *C. arabica* (Genoma C^aE^a) es un alotetraploide formado por hibridación entre dos especies diploides: *C. eugenioides* (Genoma E) como progenitor femenino, y *C. canephora* (Genoma C). El origen de la poliploidización que dio lugar a híbridos fértiles fue propuesto a partir de dos teorías: la duplicación de los cromosomas en un híbrido interespecífico diploide o el retrocruzamiento de un triploide espontáneo (Harlan y deWet 1975).

Las dos especies parentales están estrechamente relacionadas, y los dos subgéneros tienen baja divergencia de secuencia, es decir, 1,3% de diferencia promedio para los genes, (Cenci *et al*., 2012). A pesar de la estrecha relación entre los dos subgenomas constitutivos, *C. arabica* muestra un comportamiento meiótico de tipo diploide con formación bivalente. Análisis moleculares más recientes mostraron que los reordenamientos genómicos que

implicaban intercambios homeólogos[6] se produjeron en *C. arabica* y podrían ser una fuente importante de diversidad genética. La evidencia de un gran número de eventos de intercambio homeológico (HEE) compartidos por todas las accesiones de *C. arabica* apoya firmemente la hipótesis de un solo evento de alopoliploidización (Lashermes *et al*., 2014).

El tamaño del genoma del café ha sido estimado en 1,3 Gb para *C. arabica* (De Kochko *et al*., 2010) y 710 Mb para *C. canephora* (Denoeud *et al*., 2014). Mediante citometría de flujo, el contenido de ADN de *C. arabica* se estimó como 2,47 pg (equivalente a 1,2 Gb) y en *C. canephora* como 1,43 pg (equivalente a 700 Mb), mientras que en otras especies de café diploides el contenido de ADN varió de 0,96 pg (469 Mb) (*C. mauritiana*) a 1,84 pg (900 Mb) (*C. humilis*). (Sattler, Carvalho, & Clarindo, 2015b).

A pesar de la importancia económica del café, el primer borrador del genoma fue reportado para el café robusta sólo hasta el 2014 por (Denoeud et al., 2014). Para generar esta secuencia del genoma de *C. canephora*, se usaron diferentes técnicas de secuenciación, incluyendo Roche 454 de lectura sencilla y pareado, lecturas Sanger de librerías BAC-end, con una cobertura 29,5X del tamaño del genoma (710 Mb) y secuenciación Illumina®, con una cobertura 60X, la cual se utilizó para corregir errores de secuenciación y llenar los vacíos. El ensamblaje resultante tiene 25.216 *contigs* y 13.345 *scaffolds* con una longitud total de 568,6 Mb. Se obtuvo un mapa genético de alta densidad, con un total de 349 scaffolds que cubren aproximadamente 364 Mb asociados a los 11 cromosomas de *C. canephora*, entre los cuales 139 estaban asociados con orientación. Se identificaron elementos transponibles (TEs) en el genoma *C. canephora*, representando casi la mitad (49,2%) de la longitud del genoma de referencia. Igualmente, se ha analizado la transferencia de los genomas de organelos al genoma nuclear con una cantidad típica de fragmentos del cloroplasto (2014 inserciones, lo que representa el 0,16% del genoma) y un fragmento inusualmente grande de 750 kb de mitocondria, más que en otros genomas de plantas. En total, se obtuvieron 25.574 modelos de genes codificantes de proteínas en funciones implicadas en la biosíntesis de compuestos secundarios responsables del sabor del café y el aroma. Fueron examinadas familias de genes específicos para cada especie tales como las N-metiltransferasas (TMR), genes relacionados con la defensa, alcaloides y flavonoides (Dereeper *et al*., 2015).

Los estudios en diversidad genética del café en *C. arabica* y *C. canephora* han arrojado diferentes resultados. En general entre los tres principales materiales: (i) cultivares-variedades; (ii) accesiones-introgresiones-híbridos; y (iii) espontáneos-subespontáneos. Para el grupo i demostraron una baja variación genética entre los cultivares de *C. arabica* colectados de diferentes regiones tales como Brasil, Yemen, India, Australia, Vietnam, Tanzania y Hawaii, usando los diferentes marcadores moleculares como, *Random Amplified Polymorphic DNA* (RAPD), Amplified Fragment Length Polymorphism (AFLP), Simple Sequence Repeat (SSR), Inter Simple Sequence Repeat (ISSR), Sequence-Related Amplified Polymorphism (SRAP), Target Region Amplification Polymorphism (TRAP) (Motta *et al.*, 2014).

Para el grupo ii se demostró un espectro de variación bajo y significativo, con una mayor diversidad en las líneas introgresadas.

El último grupo (espontáneos-subespontáneos), en el que se esperaría la mayor diversidad, tuvo una variación que estuvo entre baja, moderada y alta. También se ha observado una correlación genética con la geografía-origen mostrado en algunos estudios. En el trabajo presentado en el informe anual de investigación mundial del café de 2014, la

[6] Cromosomas hermanos.

evaluación de la diversidad genética de 800 accesiones de café arábigo de la colección del Centro Agronómico Tropical de Investigación y Enseñanza-CATIE (Costa Rica), demostró una diversidad genética menor de esta especie en comparación con otros cultivos importantes. Mostrando que las variedades cultivadas de café contienen aproximadamente el 45% de la diversidad genética[7], indicando la baja variabilidad de la especie como una limitación para los programas de mejoramiento genético, dificultando el trabajo de mapeo genético de *C. arabica*, y por lo tanto, para la selección del germoplasma los cuales son una de las herramientas principales para el mejorador genético.

A diferencia de *C. arabica*, *C. canephora* posee una alta diversidad genética debido a su origen, al modo de reproducción y la diseminación. Se estructura en dos grupos principales (Guinea y Congo) con seis subgrupos. El desequilibrio de ligamiento evaluado en poblaciones usadas en mejoramiento de *C. canephora* es variable, lo que indica que es posible aplicar los estudios de asociación dentro de las especies *C. canephora* y utilizarlos para mejorar los de *C. arabica* debido a la relación filogenética con esta especie. La comprensión de la diversidad genética de las dos especies dominantes es crítica para cualquier esquema de mejoramiento genético en las características deseadas en el café como producción y la resistencia a enfermedades (Tran *et al*., 2016).

2.2.1.1. Origen del genoma de C. arabica

El ADN genómico marcado de *C. canephora* y *C. eugenioides* fue usado simultáneamente como sondas para el *Genome in situ Hibridization* (GISH) en *C. arabica* permitiendo discriminar dos conjuntos de 22 cromosomas cada uno. Un grupo mostró una mayor afinidad para la sonda del genoma de *C. canephora*, mientras que el segundo grupo se hibridó preferencialmente con la sonda *C. eugenioides*. Estos resultados demostraron que el genoma de *C. arabica* surgió de la asociación de cromosomas pertenecientes a estos dos genomas ancestrales diferentes.

También al comparar los patrones de RFLP de las especies diploides silvestres con los de *C. arabica*, se pudo especificar el origen de los dos conjuntos de cromosomas o genomas, combinados en *C. arabica*. De acuerdo con las relaciones filogenéticas deducidas tanto de los patrones de variación en el ADN de cloroplastos como en el ADN ribosómico nuclear (Lashermes *et al*., 1997a; Cros *et al*., 1998), las accesiones de C^aE^a diploides incluidas en este análisis se dividieron en dos grupos (es decir, tipo canephora y tipo eugenioides). Por lo tanto, estos datos apoyan la hipótesis de que *C. arabica* es el resultado de la asociación de dos genomas, a saber, C^a y E^a. Entre los taxa examinados, *C. canephora* y *C. eugenioides* aparecen como las especies existentes más similares a los genes C^a y E^a, respectivamente. Sin embargo, los límites taxonómicos entre las especies diploides que pertenecen al grupo de los canepóides son en cierta medida cuestionables, ya que estas especies son fácilmente cruzables y producen híbridos altamente fértiles (Berthaud & Charrier 1988). Cuando hay más muestras disponibles después de extensas misiones de prospección, una mayor especificación de progenitores puede quizás basarse en estudios genéticos poblacionales.

2.3. Mapeo Genético y QTLs en café.

[7] Mientras que en otros cultivos de importancia económica como el tomate el 2% de la diversidad se encuentra en el tomate cultivado.

El mapeo genético es un paso importante para los programas de fitomejoramiento. El objetivo del mapeo del genoma es ubicar una serie de marcadores moleculares en sus respectivas posiciones en el genoma. Al menos quince mapas genéticos fueron publicados y están disponibles en Coffea sp. creando una enorme base de datos para el trabajo genético. Las aplicaciones de los mapas genéticos de Coffea incluyen desde la caracterización de genes hasta el análisis genómico comparativo con diferentes especies de plantas. Actualmente, la técnica de secuenciación de alto rendimiento ha permitido construir un mapa genético más rápido y preciso permitiendo la validación del mapa genético relacionado con la predicción de QTL y genes adyacentes relacionados con rasgos de importancia agronómica para Coffea sp. (Priyono & PUTRANTO, 2016).

Crouzillat y colaboradores en 2004 desarrollaron el primer mapa consenso de robusta, un mapa genético de alta resolución que comprendió 1.258 cM de longitud con 1.7 cM de densidad. El alto número de marcadores SSR aumentó el número de polimorfismos demostrando que este tipo de marcadores era adecuado para crear un mapa consenso para estudios genéticos en todo el género. En el mismo año, Pearl et al. (2004) construyeron un mapa de 1.803 cM de longitud utilizando un gran número de marcadores AFLP (464 loci). Finalmente, con 10.2 cM de densidad de marcadores, este sexto mapa no podría considerarse como mapa de alta resolución. Sin embargo, fue el primer mapa publicado disponible para el café arábica. En el mismo año, Teixeira-Cabral (2004) realizaron otro mapa genético parcial. Para este mapa se usaron 82 marcadores RAPD, conocidos también como marcadores dominantes, para confirmar el comportamiento meiótico diploide en Coffea arabica. Sin embargo, el bajo nivel de polimorfismo en las especies de arábica seguía siendo un obstáculo para el mapeo del genoma del café.

El mapa genético número ocho fue construido por Hamon *et al*. (2005) utilizando 222 marcadores AFLP y SSR. La longitud de este mapa genético interespecífico fue de 1360 cM con 7,2 cM de densidad. Otro mapa genético intraespecífico fue desarrollado por López y Moncada (2006) en C. liberica y C. eugeniodes utilizando 76 marcadores SSR que dieron como resultado una longitud de mapa corta de 378 cM con 5.0 cM de densidad. Estos marcadores intraespecíficos se utilizaron para estudiar la reproducción asistida por marcadores para cada especie analizada. Hendre y Aggarwal (2007) construyeron un mapa de ligamiento de la planta robusta diploide utilizando una población de cruce de pseudo-prueba con 1,234 cM de longitud su construcción usó 374 marcadores para cubrir 16 grupos de enlace. Se esperó con este mapa que se identificaran QTLs para la tolerancia a la sequía. Siguiendo los estudios genéticos de las especies de Coffea, Priolli *et al*. (2008) creó un mapa genético interespecífico de C. arabica y C. canephora que comprende una longitud de 1.011 cM con una densidad de marcadores moderada (5.9 cM).

El siguiente mapa genético se derivó de C. arabica por Priolli *et al*., (2008). Este mapa genético fue el segundo para el café arábica, pero el primero saturado para la resistencia a la roya. Se generó un mapa con una longitud de 137.4 cM con una distancia de marcador baja (9.4 cM) utilizando marcadores AFLP complementados con marcadores SCAR. En términos de enfermedades de las plantas, una mayor eficiencia en la genotipificación realizada por SCAR permitió a los SCAR amplificar fragmentos de ADN relacionados con la resistencia y la susceptibilidad de C. arabica a la roya. Por lo tanto, el objetivo del estudio fue introducir marcadores de ADN para la selección asistida por marcadores para la resistencia a Hemileia vastratix, agente causal de esta enfermedad.

Un mapa genético de alta densidad más completo fue construido por la Red Internacional de Genómica del Café (ICGN por sus siglas en inglés) en colaboración con el

IRD / CIRAD, el Centro de I + D de Nestlé y el Instituto de Investigación del Café y el Cacao de Indonesia (ICCRI) en 2015. Este mapa fue la mejora de los mapas de Crouzillat et al. (2004) y Lefebvre-Pautigny *et al.*, (2010). Como se informó en la reunión de la ICGN, hubo tres etapas en el desarrollo usando miles de marcadores moleculares (RFLP, SSR, SNP) y la aplicación de la secuenciación de ADN asociada a la restricción (RADseq). La longitud del mapa final fue de 1471 cM con alta densidad de marcadores (un marcador cada 220 Kb) (Priyono & PUTRANTO, 2016).

El siguiente paso de un mapa genético es la identificación de QTLs que puedan relacionarse con genes adyacentes asociados con rasgos agronómicos y tecnológicos. Leroy *et al*., (2011) construyeron un mapa genético que comprendía 236 marcadores moleculares, realizando un mapeo de intervalos compuestos. La calidad de la bebida se evaluó en relación con los rasgos bioquímicos y de degustación de taza. Se identificaron QTLs para casi la mitad de los rasgos evaluados, con efectos que oscilaban entre el 6% y el 80% de la variación fenotípica. Los QTLs más fuertes explicaron un alto porcentaje de la variación para el rendimiento en 2006 (34% a 57%), tamaño del grano (25% a 35%), contenido de ácidos clorogénicos (22% a 35%), contenido de sacarosa y trigonelina (29% a 81%), y acidez y amargor de las bebidas de café (30% a 55%). Se identificaron regiones del genoma de C. canephora que influyen en la calidad de las bebidas. Se ubicaron cinco QTLs con genes candidatos relacionados con la biosíntesis de los rasgos analizados: dos genes que codifican la biosíntesis de cafeína, un gen implicado en la biosíntesis de ácidos clorogénicos y dos genes implicados en el metabolismo del azúcar. Este es uno de los primeros estudios sobre la identificación de QTLs que combinan rasgos agronómicos y de calidad en el café. La alta variabilidad de los rasgos de calidad dentro de C. canephora y la presencia de QTLs consistentes ofrecen a los mejoradores una herramienta prometedora para mejorar la calidad de la taza de café. Este estudio se basó en una progenie de "retrocruzamiento" entre los grupos congoleños y guineanos en los que se estableció el programa de selección en Costa de Marfil. De los resultados de este estudio, surgió como perspectiva el mejoramiento de la calidad de C. canephora (Leroy, De Bellis, Legnate, Kananura, Gonzales, Pereira, Andrade, Charmetant, Montagnon, Cubry, Marraccini, Pot, & de Kochko, 2011b)

Este trabajo anterior se ve complementado con el trabajo de Lepelley *et al*., (2012), que fue dedicado a la familia de los genes PAL de café, lo que llevó a la identificación, caracterización y mapeo de tres genes. Su expresión diferencial, y en particular la asociación establecida entre la alta expresión observada para los genes PAL1, PAL3 y Hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferasa (HQT) y la alta acumulación de ácidos clorogénicos en el grano de café inmaduro. El impacto de este estudio es el posible uso de estos marcadores en el desarrollo de variedades ricas en compuestos antioxidantes beneficiosos para la salud humana o con mejor calidad organoléptica. (Lepelley *et al*., 2012)

Gartner *et al*., 2013, construyeron un mapa genético conteniendo 146 loci codominantes y 11 grupos de ligamiento. La longitud del mapa materno / paterno unido es de 798,68 cM, tiene una densidad de saturación promedio de 6,01 cM por marcador y cubre aproximadamente el 66–86% del genoma del café diploide. Se evaluó la segregación genética en 618 loci marcadores moleculares en una población de mapeo interespecífica que constó de 101 plantas híbridas F1 de un cruce entre la especie diploide *C. liberica* y *C. eugenioides*. De estos, 168 SSR y dos EST exhibieron patrones polimórficos que permitieron el análisis de segregación y las estimaciones de enlace genético. Aproximadamente el 24% de los loci tenían alelos nulos, y el 23.5% exhibió distorsión de la segregación. Las comparaciones de los dos mapas parentales con el mapa unificado dejan en claro que, para los fines de la

reproducción genética, es crucial utilizar mapas parentales. El mapa unificado obtenido en la población F1, aunque matemáticamente robusto, no necesariamente se corresponde biológicamente con los reordenamientos cromosómicos que pueden haber surgido durante el proceso de evolución de los padres porque no son el resultado directo de la recombinación meiótica que ocurre en una población de mapeo. El fenómeno de distorsión de la segregación fue detectado en este estudio y es comparable al encontrado en otros estudios (Paillard *et al*. 1996; Ky *et al*. 1999; Lashermes *et al*. 2001; Pearl *et al.* 2004). Usando un criterio de B 0.01, se distorsionaron 40 loci, lo que corresponde a 23.5%. De estos, diez (5.9%) fueron segregantes tipo 1 (Heterocigotos en el primer padre *C. eugenioides* (P1); Permite la detección de dos o tres alelos, y los genotipos posibles son ll y lm en proporciones 1: 1. Informativo del padre P1), 19 (11.2%) fueron segregantes tipo 2 (Heterocigotos en el segundo progenitor C. liberica (P2); Permite la detección de dos o tres alelos, inverso al Tipo 1. Los genotipos posibles son nn y np en proporciones 1: 1. Informativo del padre P2), dos (1.2%) fueron segregantes tipo 3 (Heterocigotos en ambos padres; Permite la detección de dos alelos. Los posibles genotipos son hh, hk y kk, en proporciones 1: 2: 1. Cuando se incluyen alelos nulos, h0 y k0 son genotipos dominantes de h o k sobre el alelo nulo, los genotipos son h0 y kk o hh y k0 en proporciones 3: 1 o 1: 3, respectivamente. En cualquier caso, la información de segregación es parcial), y 9 (5.3%) fueron segregantes tipo 4 (Los heterocigotos en ambos padres permiten la detección de cuatro alelos sin dominancia, y los genotipos posibles son ac, ad, bc y bd en proporciones 1: 1: 1: 1. Informativo de ambos padres). Estos valores se consideran promedio. para las plantas y, dado que la población investigada se derivó de un cruce interespecífico; es muy probable que este aspecto haya desempeñado un papel importante en las distorsiones de segregación que se encuentran aquí. (Gartner, McCouch, & Moncada, 2013)

La disponibilidad de marcadores que están estrechamente vinculados a los genes de resistencia permitiría el rastreo de nuevas fuentes de resistencia y, por lo tanto, la identificación de genotipos resistentes sin la inoculación del patógeno sería de gran ayuda, ya que disminuiría la duración de los programas de reproducción. En este escenario, los marcadores moleculares son herramientas poderosas en los programas de pirámidización en los que se introducen diferentes genes de resistencia en la misma línea de cultivo de café para aumentar la efectividad de la resistencia. En general, los hallazgos presentados en el estudio de Romero *et al*., sugieren fuertemente que cuatro marcadores de SSR candidatos se asociaron con una alta resistencia abarcando una región de 2.5 cM designada como QCLR_4 ubicada dentro del cromosoma 4 del mapa internacional de *C. canephora* y que esta región QCLR_4 está involucrada en resistencia a roya observado en los datos de muchos genotipos derivados de híbrido de timor, en particular los derivados de CIFC 1343 y 832/2. Para identificar marcadores estrechamente vinculados a dichos genes, se generó una población de mapeo F2 derivada de un cruce entre 'Caturra' (variedad susceptible) y la línea DI.200 derivada de híbrido de timor (altamente resistente). Utilizando información de secuencia expresada y un enfoque de bioinformática, se identificaron los marcadores de polimorfismo amplificado de la región dirigida (TRAP) y los marcadores de repetición de secuencia simple (SSR). Las evaluaciones fenotípicas en el campo y bajo condiciones controladas confirmaron la existencia de un locus de rasgo cuantitativo para la resistencia a roya. La presencia de la región QCLR_4 se confirmó en un conjunto de líneas de élite y variedades comerciales. correspondiendo a un locus SH nuevo y genéticamente independiente que podría ser útil en el proceso de pirámidización de genes con otros genes para mejorar la resistencia a la roya en los derivados de híbido de timor. (Romero, Vásquez, Lashermes, & Herrera, 2013)

Para 2015 Pestana *et al*., realizaron análisis de marcadores moleculares determinando que la resistencia de HDT UFV 443-03 a la raza II está condicionada por al menos dos loci dominantes independientes, un locus está ubicado en el LG 2 y el otro en el LG 10. La resistencia a la raza I y el patotipo 001 están condicionadas por al menos cuatro loci dominantes independientes, dos loci están ubicados en el LG2, uno en el LG3, y uno en el LG10. Los mismos cuatro loci independientes condicionan la resistencia a la raza I y el patotipo 001. Dos de estos cuatro loci independientes también condicionan la resistencia a la raza II. Esto apoyaría la teoría que la pirámidización de genes podría resultar en un cultivar con resistencia duradera; sin embargo, es difícil distinguir entre plantas con uno o más genes de resistencia debido a efectos epistáticos. Los marcadores moleculares vinculados a los genes de estos loci de resistencia se indicaron para la selección asistida por marcadores derivados de HDT UFV 443-03, ya que es una alternativa de reproducción eficiente para la resistencia a roya sin efectos epistáticos de este tipo, debido a que la selección fenotípica para la resistencia a roya es un proceso complejo y lento en café (Pestana *et al.*, 2015).

Moncada *et al*., 2016 realizaron el mapa genético donde se identificron QTLs asociados a rasgos de interés como producción, altura de la planta y tamaño del grano, el cual es descrito con mayor profundidad en la sección de antecedentes, ya que fue utilizado para la ejecución de este trabajo.

En 2017 Akaffou *et al,* publicaron que los parámetros de fertilidad y productividad en la F1 de la población de mapeo fueron mucho más bajos que los de las especies parentales. De hecho, del 5-13% del polen era viable según el año, la tasa de fructificación fue del 9.6%, el número de semillas fue de 1.02, el peso de 100 frutos fue de 8 g y se desarrollaron siete flores por nudo. Estos parámetros no difirieron de los registrados para el híbrido BC1. Después del análisis de regresión lineal múltiple, se encontraron relaciones significativas y positivas entre la tasa de fructificación y el conjunto de semillas ($r = 0.54$) y entre el número de flores por nodo y el peso de 100 frutos ($r = 0.26$), mientras que no se calcularon correlaciones significativas entre estos cuatro caracteres (FR, SSET, NFN y W100) y la viabilidad del polen. El análisis de componentes principales reveló que los tres factores principales de variabilidad entre los descendientes BC1 eran la fertilidad femenina y masculina, así como el rendimiento. Los principales resultados de este estudio podrían resumirse en tres puntos: 1) la F1 y la mayoría de las plantas BC1 florecieron a fines de febrero, como el C. pseudozanguebariae y el progenitor silvestre; 2) los parámetros de productividad fueron muy bajos en las plantas F1 y no se restauraron totalmente en promedio en la primera generación de retrocruzamiento; y 3) se observó una alta variabilidad entre las plantas BC1, lo que permitió la selección de plantas interesantes para otros retrocruzamientos. Estos resultados fueron altamente prometedores para la mejora de Coffea canephora a través de la transferencia de genes de C. pseudozanguebariae. (Akaffou, Konate, Sié, Crop, 2014, n.d.)

Finalmente, en Ferrao *et al*., 2018, se observa un trabajo inicial donde se realiza la predicción genómica en la reproducción del café. En particular, este equipo de trabajo se centró en modelos aditivos para la selección genómica. Teniendo en cuenta que C. canephora es una especie reproducida clonalmente, los modelos predictivos diseñados para explorar el valor genético total per se de un individuo es una pregunta relevante. Para predecir los efectos aditivos se tuvo en cuenta el contexto de la reproducción, que incluye la acumulación de alelos favorables a través de ciclos tempranos y cortos de selección recurrente. Los resultados de este estudio son lo suficientemente prometedores para justificar investigaciones adicionales, incluida la extensión al modelamiento de efectos no aditivos y la incorporación

de interacciones genotipo por ambiente (G × A) al considerar el beneficio de las correlaciones genéticas entre localidades (López -Cruz et al. 2015; Cuevas et al. 2016, 2017). También se enfatizó que una de las principales dificultades para realizar estudios de café es la falta de información sobre la arquitectura genética de rasgos complejos (Tran et al. 2016). Por lo tanto, los avances en la comprensión básica de la arquitectura genética, que incluye estudios a nivel genómico y fenotípico, conducirán a mejoras adicionales en la selección genómica. (Ferrão *et al*., 2018). Algunos de los hallazgos clave de esta investigación fueron: (i) obtener capacidades predictivas similares en la mayoría de los modelos que se compararon, de manera consistente con estudios anteriores que utilizan datos reales; (ii) los modelos predictivos exhibieron costos computacionales muy diferentes; (iii) el método RR-BLUP (Regresión RIDGE – mejor predicción lineal no sesgada) logró un buen equilibrio entre capacidad predictiva y costo computacional; (iv) la diversidad y la relación genética entre los conjuntos de datos de entrenamiento y pruebas son requisitos importantes; y (v) la capacidad predictiva positiva apoya la idea de implementar selección genómica en esquemas convencionales de selección recurrente en café. En comparación con los métodos fenotípicos tradicionales, se espera que la implementación de la selección genómica pueda acelerar los ciclos de reproducción en el café (Ferrão *et al*., 2018).

La genotipificación molecular está en marcha y el análisis de QTL se podría utilizar para un esquema de selección asistido por marcadores o selección genómica, disminuyendo tiempo y costos en un programa de mejoramiento genético de una planta perenne como el café.

2.4. Antecedentes

El café es un cultivo de gran valor debido a su sabor, aroma y los efectos estimulantes de la cafeína. Denoeud *et al*., 2014, obtuvo un genoma borrador de la especie *C. canephora*, que muestra un orden de genes cromosómicos conservados entre las angiospermas asterides. Aunque no se muestra signos de la triplicación del genoma completo identificada en especies de solanáceas como el tomate, el genoma incluye varias expansiones familiares de genes específicos de la especie (Denoeud *et al*., 2014).

La diversificación funcional del genoma a través de la duplicación en tándem pudo haber dado origen a características específicas de la composición química del café en el grano. La abundancia máxima de la transcripción coincide con el aumento dramático en el contenido de ácido linoleico que se produce durante el desarrollo de la semilla en la transición perisperma-endospermo. El indicio genómico adaptativo de *C. canephora* identifica la evolución convergente de la biosíntesis de cafeína entre los linajes de las plantas y establece el café como una especie de referencia para comprender la evolución de la estructura del genoma en las angiospermas asterides (Denoeud *et al*., 2014).

Moncada *et al*., (2016), identificaron QTLs asociados con el rendimiento, la altura de la planta y el tamaño del grano. El mapa genético fue generado con una población de mapeo F2 compuesta por 278 individuos la cual fue desarrollada a partir de un cruce entre *C. arabica* var. Caturra y una accesión silvestre de *C. arabica* de Etiopía, CCC1146. Un total de 338 *Single Sequence Repeats* (SSR) se utilizaron para construir un mapa de ligamiento de trabajo. Posteriormente, se agregaron marcadores *Single Nucleotide Polymorphism* (SNP) para adicionar más marcadores al mapa y que este fuera más robusto. Se obtuvieron 22 grupos de ligamiento poblados por 848 marcadores SSR y marcadores SNP, con una longitud total del mapa de 3800 cM. Caturra es una variedad comercial de porte bajo distribuido ampliamente

en Colombia y en otros lugares de América. Sus propiedades incluyen una buena calidad de taza, con un contenido de cafeína de 1,15%, alto rendimiento, tamaño de grano pequeño y susceptibilidad a *Hemileia vastatrix* (Berk. & Broome, 1869) (roya del cafeto) y a la enfermedad de la cereza del café (*Coffee Berry Disease*-CBD) por *Colletotrichum kahawae* (Waller and Bridge, 1993). La accesión silvestre CCC1146 también tiene una buena calidad de taza, pero es contrastante en cuanto a que presenta una mayor altura, un contenido de cafeína de 1,30%, mayor rendimiento, mayor tamaño del grano y resistencia a la roya (*H. vastatrix*) y al CBD (*C. kahawae*), los cuales son rasgos de interés para el programa de mejoramiento genético del café.

Los experimentos para la asociación de rasgos a QTLs se llevaron a cabo en estaciones experimentales en cinco ambientes diferentes en Colombia. Todos los experimentos se establecieron utilizando un Diseño de Bloques Aleatorios (DBA) con dos repeticiones, y las representaciones de cada progenie consistieron en diez plantas F3 por fila para cada replica. Debido a que el gen para altura de la planta pequeña es dominante, la segregación para la altura de la planta se observó en la generación F2.

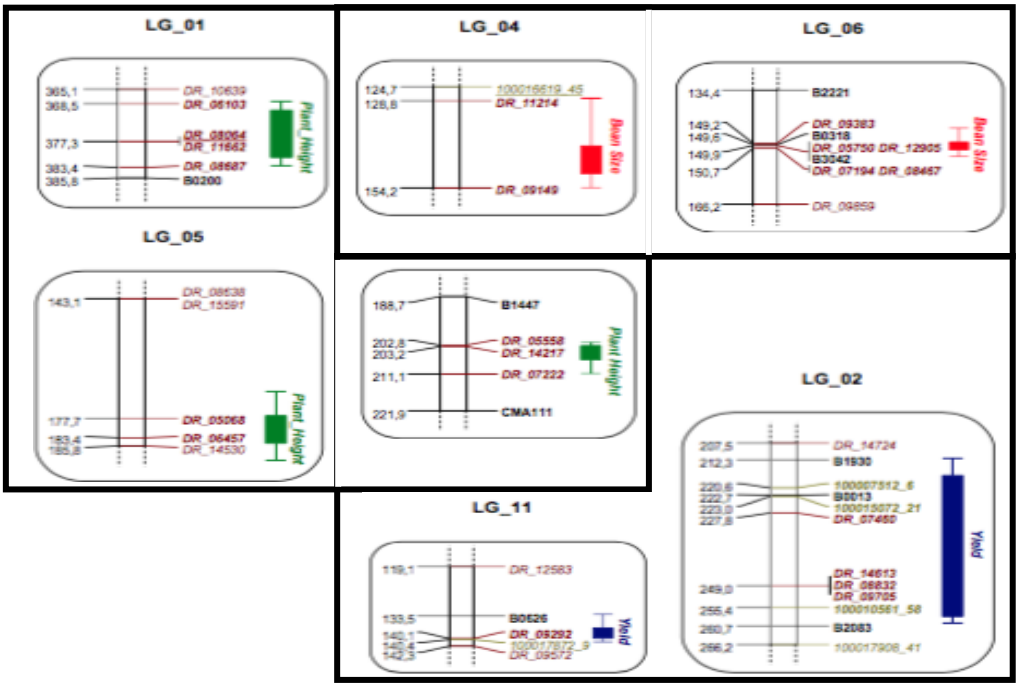


Figura 5. QTLs identificados en la población F3 para producción (*yield*), altura de la planta (*plant height*) y tamaño del grano (*bean size*).

Modificado de: (Moncada *et al*., 2016).

Para el análisis QTL, se plantaron progenies F3 de plantas F2 autopolinizadas en cinco ubicaciones. Se midieron los rasgos agronómicos en un total de 278 individuos F2 como base para la determinación del genotipo y la construcción del mapa.

Las progenies F3 de la población de mapeo se plantaron en cinco lugares y se evaluó el rendimiento, la altura de la planta y el tamaño del grano. Para producción, fue identificado

un QTL en dos localidades y un segundo fue significativo en un lugar. Se detectaron dos QTLs para la altura de la planta y el tamaño del grano respectivamente (Moncada *et al.*, 2016).

Mediante el uso del mapeo de intervalo compuesto (CIM), se identificó un QTL en el grupo de ligamiento 2, ubicado a 250,98 cM, que representó el 22,2% de la variación del rendimiento en Paraguaicito, y un QTL en 221.6 cM en el mismo grupo de ligamiento que explica el 14,35% de la variación en Rosario, también fue detectado otro QTL en el grupo de ligamiento 4, en 215 cM. En el grupo de ligación 11, se identificó un QTL adicional utilizando CIM con un pico ubicado en 155,8 cM que explicaba el 16,9% de la variación de producción en Rosario. Utilizando la estrategia del análisis de punto único, se detectó el mismo rendimiento QTL en el grupo de ligamiento 2 para Paraguaicito, Rosario, y un pico subumbral adicional para el Libano.

Así mismo se detectó un QTL para la altura de la planta en el grupo de ligamiento 1 en una región centrada en 371 cM que representaba el 25% de la variación en Paraguaicito. De manera similar, se identificó un QTL en el grupo de ligamiento 4 a 204 cM que explicaba el 15,65% de la variación en Rosario. De otro lado, fue detectado un QTL en el grupo de ligamiento 5 en 177 cM en Libano, y se detectó un pico por debajo del umbral en Naranjal. (Moncada *et al*., 2016)

Finalmente, se detectaron dos QTL para el tamaño del grano en el grupo de ligamiento 4 centrado en 128,8 cM en el Tambo y el otro en el grupo de ligamiento 6 en 147,34 cM en Rosario (Moncada *et al*., 2016).

Dentro de la estrategia del Cenicafe para conocer el genoma del café y buscar genes o genes características genómicas asociadas con rasgos de interés agonómico se considero inicialmente el mapa genético de *C. arabica* variedad Caturra y el mapa físico de la misma variedad. Este último fue construido a partir de una librería BAC (sigla en ingles para cromosoma bacteriano artificial) usando el mismo individuo usado como progenitor materno de la población segregante del mapa genético. La librería BAC y el mapa físico fueron desarrollados con el apoyo del laboratorio del Dr. Rod A. Wing, Arizona Genomics Institute – AGI, de The University of Arizona. A partir de la secuenciación de los extremos de los insertos de los clones BAC o BES (sigla en ingles para BAC End Sequence) se generaron marcadores moleculares del tipo microsatelite (SSR) y SNPs que sirvieron para la construcción del mapa genético. Posteriormente, en el marco de un proyecto de colaboración entre Cenicafe y Cornell University, y contando con financiamiento del Banco Interamericano para el Desarrollo – BID, Fontagro y la National Science Foundation – NSF (USA), se secuenció el genoma de *C. arabica* variedad Caturra, del mismo individuo usado para el mapa genético y físico. Se probaron multiples metodologías de secuenciación incluyendo secuenciación de tercera generación desarrollada por la empresa PacBio, la cual produjo la menor fragmentación del genoma como producto de ensamblaje de secuencias. (Maldonado 2017, comunicación personal).

Después de obtener el mapa genético y desarrollar análisis de QTLs, como se describió previamente, Maldonado, Padilla y Wing (2014, datos no publicados) diseñaron y ejecutaron, en el Arizona Genomics Institute, un proyecto que buscaba convertir en secuencia genomica los intervalos de confianza marcados por mapeo de QTLs mediante secuenciación de regiones específicas del genoma o "*targeted sequencing*" a partir de información suministrada por Moncada (2014, datos no publicados). Sobre el mapa genético se identificaron 22 regiones con QTLs asociados con rasgos que incluyeron producción, tamaño de grano, tamaño de planta, numero de ramas, diámetro de tallo, número de lóculos por

cereza, resistencia a la roya del café (*Hemileia vastatrix*), características de grano como contenido de lípidos, trigonellina y cafeina, y características de tasa como acidez, dulzor, cuerpo y aroma, entre otros. Con toda esta información y contando con la secuencia genómica del diploide *C. canephora* (Denoeud *et al.*, 2014) se integró, mediante herramientas bioinformáticas, la secuencia y los mapas genético y físico de *C. arabica*. Para la estragia de "*targeted sequencing"* a partir del mapa físico se seleccionaron los contigs que cubrían regiones de 2 a 3 Mb corriente arriba y abajo de la posición de los marcadores asociados a QTLs sobre el genoma de *C. canephora*; posteriormente, para cada uno de los contigs del mapa físico se seleccionó el minimo numero de clones BAC con un sobrelape de al menos 30 bandas de "*fingerprint"* o MTP (sigla en inglés para Minimum Tilling Path), para ello se utiliaron los programas FPC (Soderlund, Longden, & Bioinformatics, 1997) y SyMap (Soderlund, C., Nelson, W., Shoemaker, A., & Paterson, A. 2006), forzando la inclusión en el MTP de los clones que contenían en sus secuencias BES los marcadores moleculares incluidos en el mapa genético. Seleccionados los clones BAC, clonados en *Escherichia coli*, se cultivaron en medios de cultivo liquido específicos y se purificó el ADN de cada clon BAC, los que se secuenciaron por medio de tecnología PacBio siguiendo una estrtegia de construcción de librerías de secuenciación de grupos de clones BAC que no fueran contiguos según el mapa físico. La secuencia producida se ensambló usando el programa HGAP desarrollado por PacBio y la circularizacion de los clones BAC seguido de limpieza de secuencia de vector y asignación de identidad del clon BAC mediante alineamiento secuencias BES se hizo usando un programa llamado Post-HGAP desarrollado en el AGIy descrita en Zhang y colaboradores (Zhang *et al*., 2016) quienes secuenciaron dos genomas de arroz siguiendo la estrategia BAC por BAC a partir de maspas físicos. Con el propósito de determinar polimorfismos entre los progenitores de la población usada para el desarrollo del mapa genético y mapeo de QTLs, en regiones genómicas codificantes e intergenicas, se secuenció el genoma del progenitor masculino a baja cobertura mediante tecnología Illumina (~15X, PE 2x150) y se secuenciaron genes completos de ambos progenitores mediate tecnología Iso-Seq de PacBio siguiendo un método de agrupamiento de ARNm marcado diferencialmente por tejido con etiquetas moleculares que fue desarrollado en el AGI por Maldonado y Padilla en el 2014 contando con el apoyo del laboratorio de desarrollo de la empresa PacBio. (Maldonado 2017, comunicación personal).

Tras la publicación del mapa genético de *C. arabica* junto con QTLs para producción, tamaño de grano y atura de la planta de café por Moncada y colaboradores (2016) se suministro para este trabajo parte de la información descrita en los párrafos anteriores, además de secuencias sensabladas del genoma de *C. arabica* var. Caturra, que hacen parte de la investigacion del Programa de Mejoramiento Genetico de Cenicafé llamada "Desarrollo de herramientas para selección asistida por marcadores moleculares" (MEG106001) liderada por Maldonado (Maldonado 2015, documento interno Cenicafé).

3. Materiales y métodos

3.1. Secuencias de genoma y transcriptoma usadas en el trabajo

3.1.1. Targeted sequencing BAC por BAC Secuencias PacBio®

Se utilizaron las secuencias ensambladas provenientes de un trabajo anterior realizado en CENICAFÉ (brevemente descrito en la sección 2.4 Antecedentes) de clones de librerías BAC de *C. arabica* variedad Caturra (progenitor femenino de la población de mapeo) con una profundidad 10X, que componen el MTP de entre 4 y 6 Mpb corriente arriba y abajo a partir de los marcadores QTL asociados con los rasgos de interés. Estas secuencias fueron seleccionadas con el fin de terminar el ensamblaje de las regiones asociadas con producción, altura de la planta y tamaño del grano de acuerdo con el orden de los clones dentro del MTP (Tabla 1). (Maldonado *et al*, 2017).

3.1.2. Secuenciación del genoma completo

Se contó con secuencias de ADN de genoma completo (WGS por sus siglas en inglés) previamente ensambladas con HGAP de los parentales de la población de mapeo *C. arabica* var. Caturra con la tecnología PacBio con una profundidad de 60X, un N50 de 1´015.410 pb en 2978 *contigs*. Se usaron en total 40´871.928 lecturas de la primera línea y 39´985.208 lecturas de la segunda línea de secuenciación (Tabla 1).

Tabla 1. Resumen de secuencias que fueron usadas en el estudio.

Estrategia Secuenciación	Tecnología	Máquina	Variedad/ Introducción	Cobertura	Ensamblador	Comparación genoma de referencia
WGS	PacBio®	RS-II™	Caturra	60 X	Falcon	Symap (BLAT - BLAST)
	Illumina®	HiSeq™-2500	CCC1146	10 X	ND	Bowtie2*
Targeted sequencing (BAC-por-BAC)	PacBio®	RS-II™	Caturra	85 – 450 X	HGAP - AMOS - MUMMER	Symap (BLAT - BLAST)

ND. No se utilizaron secuencias preensambladas, los datos pareados se recibieron filtrados por calidad.
* Mapeo realizado en este estudio.

3.1.3. Librerías de secuencias completas de Genes (Iso-seq™ – PacBio®)

Se contó con 259.605 secuencias provenientes de librerías Iso-seq™ de tamaños de 1-2 kb, 2-3 kb y 3-6kb, de mRNA completo no quimérico de los genes de los parentales secuenciados *C. arabica* var. Caturra de tejidos de meristemos, hojas y flores, las cuales serían usadas posteriormente en este flujo de trabajo para predecir los genes (Tabla 2).

Tabla 2. Secuencias pertenecientes a las diferentes librerías en los parentales de la población del mapa genético de café.

Características de las librerías	Caturra			
	1-2 kb	2-3 kb	3-6 kb	Total
No. Isomorfos*	80.142	47591	17885	145618

* Genes de longitud completa no quiméricos ensamblados por HGAP y ToFU (RS_IsoSeq)

Con la información anteriormente relacionada se propuso el flujo de trabajo descrito en la figura 6, en el cual existe una primera fase de ensamblaje de la información obtenida e identificada mediante comparaciones con el genoma de referencia *C. canephora*, utilizando las secuencias de la estrategia BAC por BAC y la secuenciación WGS. Luego de la revisión de dichos ensamblajes, se realizó la anotación estructural y funcional de los genes de cada una de las regiones para altura de la planta, producción y tamaño del grano. Se determinaron las secuencias repetitivas y se compararon con las regiones de *C. canephora*. Finalmente, se identificaron las variantes entre cada uno de los subgenomas, para determinar cuales que se puedan asociar con las regiones de interés en un subgenoma específico (Figura 6).

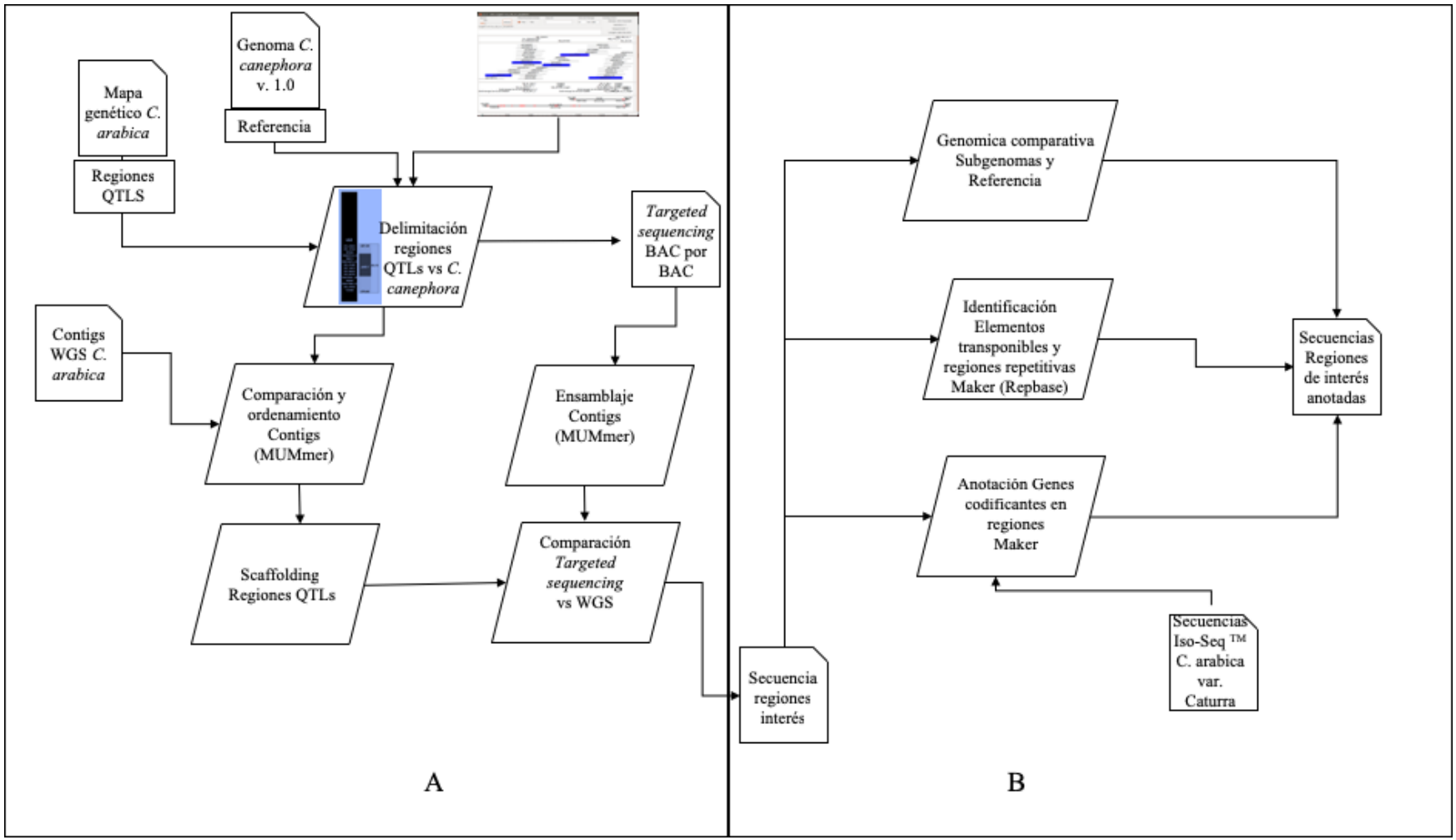


Figura 6. Flujo de trabajo para el análisis de las regiones QTL asociadas a producción, tamaño del grano y altura de la planta. (A). Actividades del primer objetivo específico. (B). Actividades segundo objetivo específico.

3.2. Identificación de las secuencias de regiones que contienen QTLs de importancia agronómica en café.

3.2.1. Delimitación de las secuencias de regiones de interés del ensamblaje WGS

El mapa genético de *C. arabica* contiene 22 grupos de ligamiento (gl), donde se reportó asociación en 3 gl para el rasgo tamaño de la planta, 2 gl para el tamaño de grano, y 2 gl para producción. Como en la publicación en la cual se basó este trabajo *(*Moncada *et al*., 2016), no se describe el intervalo de confianza de los marcadores en el mapa genético de *C. arabica* de los QTLs que mostraron correlación a la región de interés como producción, tamaño de grano y altura de la planta, se tomaron aquellos que estuvieran ubicados dentro del rango de 30 cM corriente arriba y corriente abajo de los marcadores asociados con el rasgo. Estos marcadores fueron mapeados contra las secuencias de los cromosomas de *C. canephora*, BACs y *contigs* del genoma de *C. arabica*, usando el programa BLAT (Kent, 2002) para los marcadores SNVs y DArT y el programa e-PCR (Schuler, 1997) para los microsatélites (SSR), y así obtener las secuencias contiguas (*contigs*) que pertenecen a la región.

3.2.2. Ensamblaje de datos targeted sequencing de regiones QTLs de Tamaño del grano, producción y altura de la planta.

Con el objetivo de obtener las secuencias de las regiones de interés, se tomaron los *contigs* que fueron identificados dentro de la región de 30 cM corriente arriba y corriente abajo, y se realizó el ensamblaje de acuerdo con la estrategia de secuenciación utilizada.

3.2.2.1. Ensamblaje de secuencias provenientes de WGS PacBio de Coffea arabica.

Una vez determinadas las secuencias dentro de las regiones de interés, se realizó una primera comparación y ordenamiento de los *contigs* del ensamblaje de secuenciación genómica por PacBio con MUMMer (Kurtz *et al*., 2004), usando como guía los cromosomas respectivos del genoma de referencia *C. canephora* (Denoeud *et al.*, 2014). Se seleccionaron los *contigs* correspondientes a los subgenomas de *C. arabica* (E^a y C^a), para su posterior *scaffolding*. Los subgenomas fueron seleccionados a partir de los siguientes criterios:

i) En el caso de encontrar *contigs* que cubrían simultáneamente la misma región del genoma (sobrelapamiento) de referencia se asignó al subgenoma *C. canephora* (C^a) aquella secuencia que presentaba la mayor identidad promedio.

ii) El *contigs* en sobrelapamiento de identidad promedio menor se asignaó al subgenoma *C. eugenioides* (E^a).

iii) Los *contigs* homologos a *C. canephora* que cubrían la secuencia de referencia sin sobrelapamiento se incluyeron en el proceso de *scaffolding* para los dos subgenomas.

Para cada uno de los resultados del alineamiento se determinó la identidad promedio para lo cual se utilizaron los siguientes comandos de *show-coords* usando una de las salidas del Mummer:

```
1 $ show-coords -r -g -d -c -T -o BS4.fil.delta >alignmentcoordsBS4.txt
```

Se obtuvieron las listas de los *reads* que componían el alineamiento, y se calculó el promedio de identidad:

```
1 $ awk -v IFS='\t' '{print $13}' alignmentcoordsBS4.txt | sed '/^$/d' | sed '1d' | sort | uniq > temp1.txt

2 $ while IFS="\t" read -r line; do awk -v l="$line" '{if ($13 == l) print $0}' alignmentcoordsBS4.txt >> alignmentBS4.$line.txt; done < "temp.txt"

3 $ for file in alignmentBS4.*.txt; do echo $file; awk -v IFS='\t' '{sum+=$9} END {print sum}' $file; done

4 $ for file in alignmentBS4.*.txt; do printf $file"\t" >>$file.PI.out; awk '{sum += $7 } END { if (NR > 0) printf sum / NR "\n"}' $file >>$file.PI.out;done

5 $ for file in alignmentBS4.*.txt; do printf $file"\t"; awk '{ sum += $7 } END { if (NR > 0) printf sum / NR "\n"}' $file; awk -v IFS='\t' '{sum+=$9} END {print sum}'; done
```

Posteriormente el proceso de *scaffolding* se realizó con la siguiente línea de código, para cada una de las regiones de interés (producción, tamaño del grano y altura de la planta).

```
1 $ perl -p -e '$n = 100 ; s/^>.+/'N' x $n/e' <archivo_input>| sed '1d' | grep -v '^>' | grep '^.' | tr -d '[:blank:]' | cat <( echo '>Pseudo_QTL_name') - > <archivo_output>
```

Complementado por el siguiente comando presente en FASTX Toolkit ver. 0.0.13 para que en el formato fasta el campo de las secuencias quedara en líneas de 60 caracteres.

```
1    $ fasta_formatter [-h] [-i INFILE] [-o OUTFILE] [-w 60]
```

3.2.2.2. Ensamblaje de las secuencias BAC por BAC de regiones de Interés.

De acuerdo al ordenamiento de los *contigs* FPC que conformaban el MTP de las regiones producción del grupo de ligamiento 2 (YLD2) y altura de la planta de la región de ligamiento 5 (PH5), se realizó el ensamblaje con la herramienta Minimus (Sommer, Delcher, Salzberg, & Pop, 2007). Finalmente, cuando no se presentó sobrelapamiento entre clones contiguos del MTP se incluyó un *gap* de 100 Ns (código IUPAC), usando la misma línea de código para el *scaffolding* de la secuenciación WGS para obtener la región completa.

3.3. Caracterización de los genes contenidos en QTLs de importancia agronómica en café, con énfasis en los genes predichos y expresados.

3.3.1. Anotación de genes codificantes

La predicción de los genes modelo se realizó mediante el uso del programa MAKER (Campbell *et al*., 2014). El entrenamiento del *software* para mejorar la predicción de los genes fue realizado con secuencias de longitud completa del transcriptoma de meristemos, hojas y flores de *C. arabica* var. Caturra, parental materno de la población del mapa genético, que se obtuvieron por medio de la metodología Iso-seq™ de PacBio® (Tabla 2). Para los predictores SNAP y Augustus se generaron los modelos a ser usados en la predicción *ab initi*o, de la anotación curada manualmente del genoma de *C. canephora* ver. 1.0. Además, fueron incluidas como evidencia para la generación de modelos de genes las secuencias del set de datos de isoseq descritos anteriormente.

La anotación funcional se realizó con el programa InterProScan ver. 5.24 *(Jones et al.,* 2014) usando las opciones -iprlookup -goterms y –pa, para obtener los términos GO asociados a las regiones de interés. Para la descripción de las funciones de los genes, se extrajo de la salida del InterProScan las funciones de los genes de acuerdo con la descripción de la base de datos PFAM.

3.3.2. Identificación de transposones, regiones repetitivas y genes relacionados con elementos transponibles

Se realizó la identificación de secuencias repetitivas en las regiones blanco mediante el *software* RepeatMasker (Tarailo-Graovac & Chen, 2002) usando la base de datos RepBase (Bao, Kojima, & Kohany, 2015) utilizado en el flujo de trabajo del software MAKER.

Luego, se identificaron los genes relacionados con elementos transponibles seleccionando las proteínas anotadas que contenían "transposón", "transposasa", "integrasa" o "transcriptasa inversa" en la base de datos Pfam (Finn *et al*., 2016) provenientes de la salida del InterProScan.

3.3.3. Genómica comparativa

Se realizó la comparación de cada una de las secuencias de los subgenomas asociadas a regiones de rasgos cuantitativos correspondientes a producción, tamaño del fruto y altura de la planta, utilizando el programa Mauve 2015-02-25 (Darling, Mau, Blattner, & Perna, 2004; Rissman *et al*., 2009) y mediante *dot plots* producidos por el programa Gepard (Krumsiek, Arnold, & Rattei, 2007), con el fin de observar si existían diferencias a nivel estructural en cuanto a rearreglos de las regiones comparando los subgenomas y a estos con el genoma de referencia *C. canephora*.

4. Resultados y discusión

4.1. Delimitación de las regiones de interés

Fueron identificadas las regiones flanqueantes 30 cM corriente arriba y corriente abajo de los marcadores que mostraron correlación con los rasgos asociados a tamaño del grano, altura de la planta y producción (Figura 7).

A

QTL name	LG	QTL marker	POS cM	Location	LR/LOD
yld2.1	2	100007512_6/*B0013	221.6	Rosario	27.48
yld2.1	2	100007512_6	254.9	Paraguaicito	41.93
yld2.1	2	100007512_6	221	Libano	2,6
yld2.1	2	DR_09705/DR_08832	250.98	Paraguaicito	30.41
yld4.1	4	100007512_6	215	Rosario	50.74
yld11.1	11	DR_15370/DR_12117	155.79	Rosario	21.79
ph1.1	1	DR_06103/DR_10639	371.46	Paraguaicito	28.35
ph1.1	1	DR_06103	376.15	Paraguaicito	4.78
ph4.1	4	DR_14217/DR_05558	203.23	Rosario	22.64
ph4.1	4	DR_14217	201.04	Rosario	3.6
ph5.1	5	DR_05068	177.73	Libano	10.6
ph5.1	5	DR_06457	220	Naranjal	22.38
bs4.1	4	DR_11214	128.8	Tambo	4.21
bs6.1	6	DR_09383	147	Rosario	4.1

B

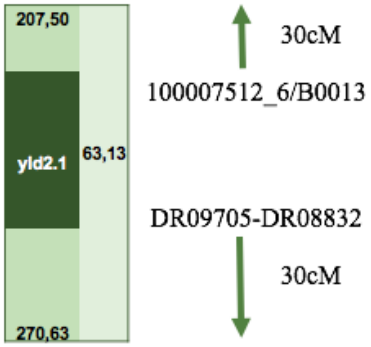


Figura 7. Ejemplo delimitación regiones de acuerdo a los marcadores moleculares asociados a producción, altura de la planta y tamaño del grano. (A). Marcadores dentro de caja verde asociados con producción*. (B). Selección región 30 cM corriente arriba y corriente abajo de acuerdo a los marcadores incluidos dentro de este intervalo.
*Modificado de (Moncada *et al*., 2016).

Es así como para el rasgo tamaño del grano (BS), BS4 fueron localizados en el genoma usado como referencia los marcadores DR_11998 y DR_05347 y para BS6 fueron identificados C444 y DR9859. Este último rasgo se eliminó del estudio porque la región que comprendía 30 cM corriente arriba y corriente abajo abarcaba 33 Mb, el cual es el tamaño de un cromosoma completo en *C. canephora*. Para el rasgo altura de la planta (PH), PH1 fueron identificados C273-2 y DR_11009, para PH4, DR_11916 y B2994, así como para PH5, DR_08638 y DR10122. Finalmente, para el rasgo de producción (YLD) los marcadores identificados fueron DR05804 y C4368 en el caso de YLD2, y DR_15516 y DR_09149 para el caso del YLD4. Este rasgo fue excluido del estudio por tratarse de una región que presentó *hits* con el cromosoma 0, es decir, presentó aciertos con las secuencias contenidas en secuencias ensambladas no ancladas en cromosomas del genoma de *C. canephora* publicado por Denoeud y colaboradores (2014). Para estas secuencias no se pudo establecer su ubicación en el genoma con base a alguna evidencia como la presencia de marcadores o mediante el uso de biblotecas mate pair según los autores (Denoeud *et al*., 2014). También se detectó que el marcador que se comparte para las regiones YLD2 y YLD4 (100007512_6) mapeó muy bien para el primero en el grupo de ligamiento adecuado (gl2), sin embargo, para el caso de la región YLD4 no presentó ningún *hit* y aquellos marcadores con un único acierto presentes en el mapa para la región, presentaron *hits* contra diferentes grupos de ligamiento

(gl 0, 11 y 8) al grupo de ligamiento cuatro que era el esperado. Además, YLD4 presentó un sobrelapamiento con la región BS4, la cual presentaba *hits* con marcadores únicos de otros grupos de ligamiento, haciendo la región inconsistente (Figura 8).

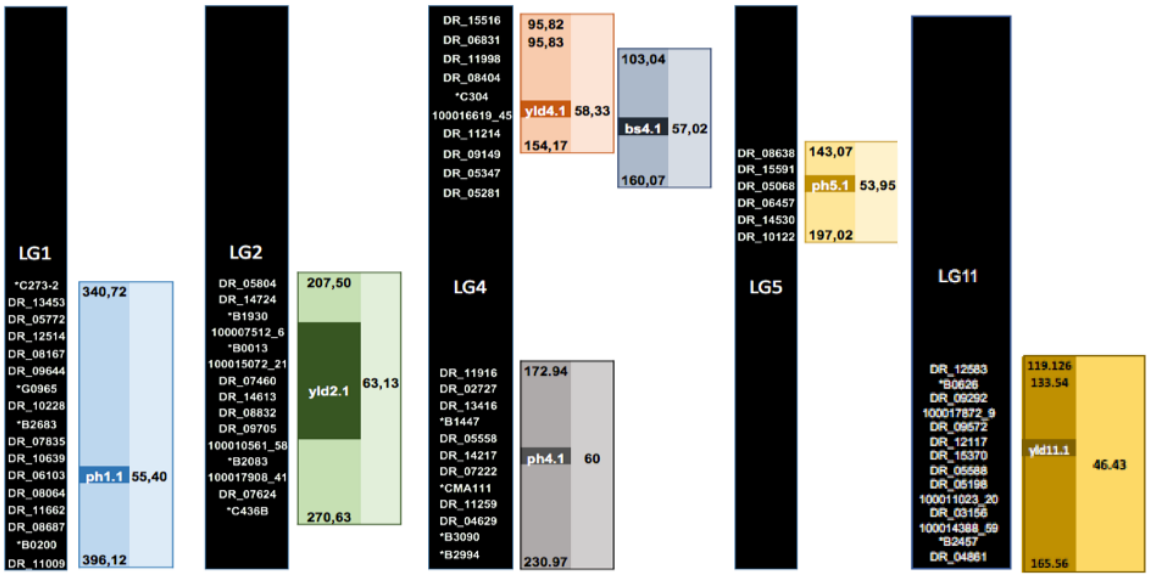


Figura 8. Delimitación de las regiones de interés en los grupos de ligamiento para los marcadores asociados a producción, tamaño del grano y altura de la planta.

Al delimitarse las regiones de interés, los *contigs* WGS de *C. arabica,* homologas a la región de interés, mostraron un patrón en cuanto al tamaño final de los mismos. Al sumarse la longitud en pares de bases antes del proceso de ensamblaje o *scaffolding*, las regiones YLD2 y PH1, presentaron un tamaño alrededor del doble de la región equivalente en *C. canephora*, convirtiéndose en un indicio de la presencia de dos secuencias por región, lo cual puede observarse en la Figura 9-A.

Este resultado se debe probablemente a la detección de más de un *contig* mapeado a la misma región y que se sobrelapan en un alto porcentaje. Es así como para la región de producción YLD2, por ejemplo, el tamaño en *C. canephora* fue de 8,66 Mb mientras en *C. arabica* fue de 15,59 Mb (Tabla 3). Una región que no exhibe ese comportamiento es la región asociada a tamaño del grano BS6, la cual tiene un tamaño de 33,93 Mb en *C. canephora* y 27,51 Mb en *C. arabica*. Como se había explicado anteriormente, el tamaño de la región BS6 equivale al tamaño promedio de un cromosoma del genoma de referencia *C. canephora* (tamaño promedio 33,1 Mb; máx. cromosoma 2- 54,5 Mb; mín. cromosoma 9- 22,3 Mb, (Denoeud *et al*., 2014) (Tabla 3).

La presencia de cromosomas homeologos en el alotetraploide, que comparten un ancestro común con el genoma de *C. canephora,* usado como secuencia de referencia, explica el por que se encuentra en estas regiones el doble de la secuencia total en *C. arabica*. Por esta razón, las asignaciones de los *contigs* se realizaron con el supuesto de que la identidad de dos *contigs* sobrelapantes con la misma región podrían ser asignados de acuerdo con la identidad promedio de la región con el genoma de referencia *C. canephora*.

Tabla 3. Tamaño de Regiones de interés en *C. arabica* y *C. canephora WGS*.

	C. arabica			*C. canephora*			
Rasgo	**gl**	**No. Ctgs**	**Len. Mb**	**Chr.**	**Start**	**End**	**Len. Mb.**
Producción (YLD2)	2	13	15.59	5	17.48	26.14	8.66
Producción (YLD11)	11	1	4.33	1	0.29	4.44	4.15
Tamaño del Grano (BS4)	4	7	6.53	8	21.16	25.61	4.45
Tamaño del grano (BS6)	6	23	27.51	4	14.62	48.54	33.93
Altura de la planta (PH1)	1	11	11.46	2	22.23	28.73	6.5
Altura de la planta (PH4)	4	7	5.75	8	5.83	12.18	6.36
Altura de la planta (PH5)	5	8	13.46	6	9.06	18.94	9.88

*Región excluida del estudio debido al tamaño de la región, el cual abarca un cromosoma completo.

Es así como aquellos *contigs* mapeados a la misma región y que presentaron una mayor identidad entre estos a la referencia, se asignaron al subgenoma C^A, y aquellos *contigs* con una identidad menor, pertenecieron al subgenoma E^A. Como se puede observar en la Figura 9-B, la asignación de dos *contigs* en un fragmento de la región YLD2 fue realizada de acuerdo con la identidad de los *contigs* con la región homóloga en el cromosoma 5 de *C. canephora*, donde el color rojo significa una identidad cercana al 100% y el color naranja denota una identidad de ~93%. Sin embargo, dentro de los *contigs* mapeados, se presentaron casos donde los *contigs* tuvieron un solo representante en la comparación, y, por tanto, estos fueron asignados para los dos subgenomas en el proceso de *scaffolding*.

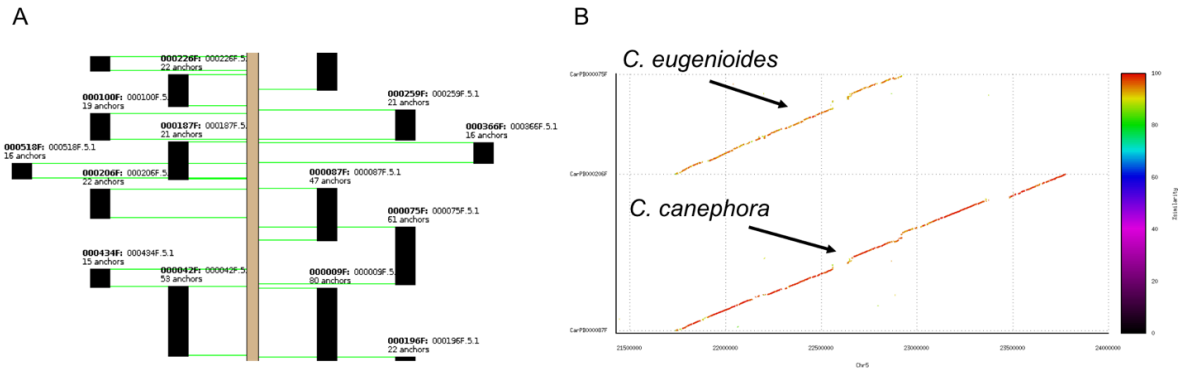


Figura 9. Mapeo de *contigs* WGS de la región YLD2 contra el cromosoma 5 de *C. canephora*. (A). Salida del SyMAP. (B). Subregión del rasgo YLD2 y asignación de *contigs* a los subgenomas de C. arabica de acuerdo con la identidad, mayor identidad (Rojo) asignada a C[a] y Menor identidad (Rojo claro con amarillo) es asignado a E[a].

4.2. Ensamblaje de datos *targeted sequencing* de las regiones QTL de tamaño del grano, producción y altura de la planta

Se asignaron los *contigs* a los subgenomas de *C. arabica C[a]* y *E[a]* de acuerdo con los resultados del BLAT y del e-PCR. Estos fueron ordenados de acuerdo con la región del cromosoma de *C. canephora* correspondiente a cada una de las regiones, para así corregir su orientación (Tabla 4).

Los marcadores contenidos a 30 cM corriente arriba y abajo del marcador DR_11214 con correlación con la característica de interés tamaño de grano (BS4) mapearon con el cromosoma 8 de *C. canephora*. Al ser comparado con el *draft* del genoma de *C. arabica*, los *contigs* que tuvieron *hits* fueron, 000377F, 000484F, 000093F, 000154F, 000208F y 000548F.

Tabla 4. Organización de los *contigs* de las regiónes de interés y asignación a los subgenomas.

QTL	Contig C[a]	COORDENADAS C[a]		DIRECCIÓN	TAMAÑO REGIÓN EN C[a] (BP)	E[a]	COORDENADAS E[a]		DIRECCIÓN	TAMAÑO REGIÓN EN E[a] (BP)
BS4	000208F	1	1249.527	+	3996426	000548F	1	598.684	+	3617251
	000093F	1.249.629	3.164.543	-		000154F	598.785	2.095.919	-	
	000377F	3.164.645	3.996.426	-		000484F	2.096.020	2.785.369	+	
						000377F	2.785.470	3.617.251	-	
PH1	000338F	1	906.255	-	6922807	000050F	1	2.583.227	+	8062471
	000390F	906.356	1.722.103	+		000158F	2.583.328	4.079.867	-	
	000715F	1.722.204	2.183.045	+		000128F	4.079.968	5.721.737	+	
	Pseudo_PH1eug	2.183.146	4.108.836	+*		000236F	5.721.838	6.884.541	+	
	000428F	4.108.937	4.872.299	-		000224F	6.884.642	8.062.471	+	
	000361F	4.872.400	5.744.877	+						
	000224F	5.744.978	6.922.807	+						
PH4	000016F	1	3.675.725	-	5502700	000143F	1	1.573.559	+	6425295
	000106F	3.675.826	5.502.700	-		000493F	1.573.660	2.244.453	+	
						000365F	2.244.554	3.110.885	+	
						000025F	3.110.986	6.425.295	-	
PH5	000140F	1	1.568.350	-	10185576	000023F	1	3.349.993	+	11894192
	000161F	1.568.451	3.032.795	-		000558F	3.350.094	3.944.812	-	
	000060F	3.032.896	5.400.136	+		000337F	3.944.913	4.864.383	-	
	000575F	5.400.237	5.975.777	-		000026F	4.864.484	8.111.662	-	
	000109F	5.975.878	7.757.749	+		000043F	8.111.763	10.798.387	-	
	000053F	7.757.850	10.185.576	-		000251F	10.798.488	11.894.192	-	

YLD2	000226F	1	1.401.207	+	10218712	000226F	1	1.401.207	+	10713067
	000259F	1.401.308	2.472.002	-		000100F	1.401.308	3.223.534	+	
	000366F	2.472.103	3.339.099	-		000187F	3.223.635	4.545.172	-	
	000518F	3.339.200	3.982.157	-		000206F	4.545.273	5.810.574	+	
	000087F	3.982.258	5.973.279	+		000075F	5.810.675	8.024.982	-	
	000075F	5.973.380	7.530.626	+ *		000042F	8.025.983	10.713.067	-	
	000042F	7.530.727	10.218.712	-						
YLD11	000343F	1	909.549	+	4528679	000343F	1	909.549	+	4774078
	000547F	909.650	1.522.385	-		000547F	909.650	1.522.385	-	
	000119F	1.522.486	3.217.033	+		000186F	1.522.486	2.834.031	+	
	000186F	3.217.134	4.528..679	+		000295F	2.834.132	3.835.761	+	
						000323F	3.835.862	4.774.078	+	

El marcador DR_06013 asociado a la característica de interés altura de la planta (PH1) fué mapeado con el cromosoma 2 de *C. canephora*; en el rasgo altura de la planta (PH4), el marcador DR_1417 mapeó con el cromosoma 8; los marcadores DR_05068 y DR_06457 con correlación a la característica de interés altura de la planta (PH5) fueron mapeados con el cromosoma 6; para el marcador DR_15370 con correlación al rasgo de producción (YLD11) fueron mapeados con el cromosoma 1 y finalmete, la región YLD2.1, los marcadores 100007512_6 y DR_09705 mapearon en el cromosoma 5 de *C. canephora*. Los contigs de *C. arabica* correspondientes a cada una de las regiones se pueden observar en la tabla 4.

Luego de la identificación de los *contigs*, la corrección en la orientación de estos y de llevar a cabo el proceso de ensamblaje y *scaffolding*, se esperaba un tamaño de 6 Mb de las regiones asociadas a rasgos de interés de acuerdo con el tamaño esperado al compararse con las regiones secuenciadas por *targeted sequence BAC por BAC* (YLD2 y PH5), sin embargo, se obtuvieron regiones más grandes de acuerdo con el mapeo de los marcadores en el genoma de referencia *C. canephora*. Esto se debe posiblemente a que las secuencias de los subgenomas dependen del tamaño de los *contigs* los cuales no fueron supeditados hasta la región del marcador que flanquea la región, como sí era el caso de la región de *C. canephora*. Lo mismo aplicó para aquellas secuencias que son más cortas (Tabla 4).

Los *dot plots* para cada uno de los *scaffolds* de los subgenomas al ser comparados con la región equivalente de *C. canephora* mostró contigüidad (Datos no mostrados). Al comparar las secuencias de los subgenomas, sólo dos compartieron las regiones al extremo 3´ (Alta similitud al final de las secuencias, observándose color rojo indicativo de similitud cercano al 100%) de la secuencia BS4 y PH1 (Figura 10 A-B). Dos compartieron regiones al inicio y al final de la región, YLD2 y YLD11 (Figura 10 E-F). Las secuencias de las regiones PH5 y PH4 (Figura 10 C-D) mostraron una menor similitud a lo largo de la secuencia ensamblada. Estas pequeñas zonas de divergencia al inicio de la región y donde pueden observarse las diferencias en identidades de los dos subgenomas son importantes debido a las múltiples aplicaciones que podrían tener, como la identificación del subgenoma al que pertenece un cromosoma en café (Figura 10).

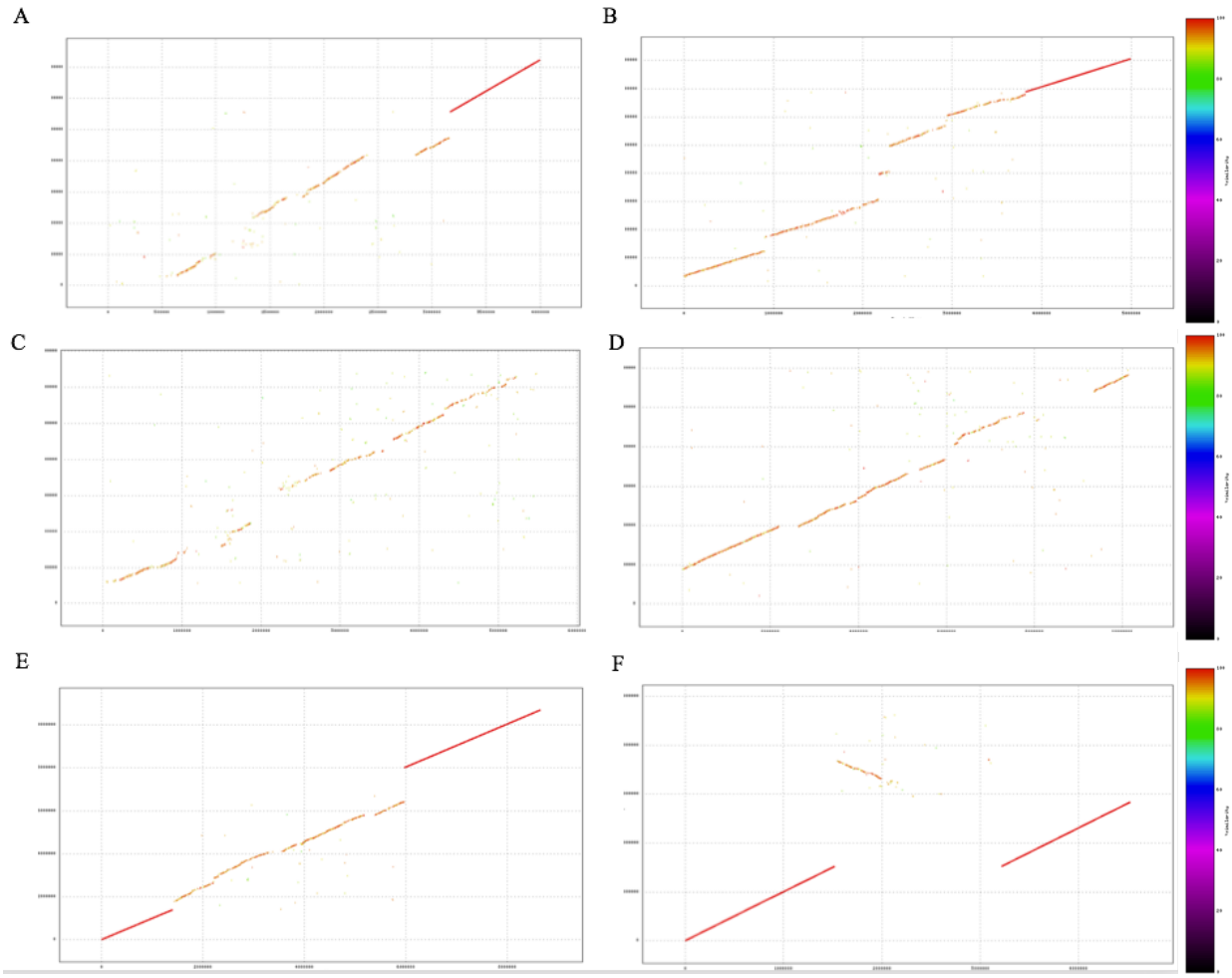


Figura 10. D*ot plots* de los subgenomas C[a] (Eje Horizontal) Vs. E[a] (Eje Vertical). Para las regiones A. BS4. B. PH1. C. PH4. D. PH5. E. YLD2. F. YLD11.

4.2.1. Ensamblaje BAC por BAC.

Los *Minimum Tiling Paths* de las regiones de interés PH5 y YLD2 fueron ensamblados. En el caso de YLD2 fueron secuenciados 126 clones BACs para una longitud inicial de 26.75 Mb (Figura 11), y para el caso de PH5, fueron 166 BACs con una longitud inicial de 33,2 Mb.

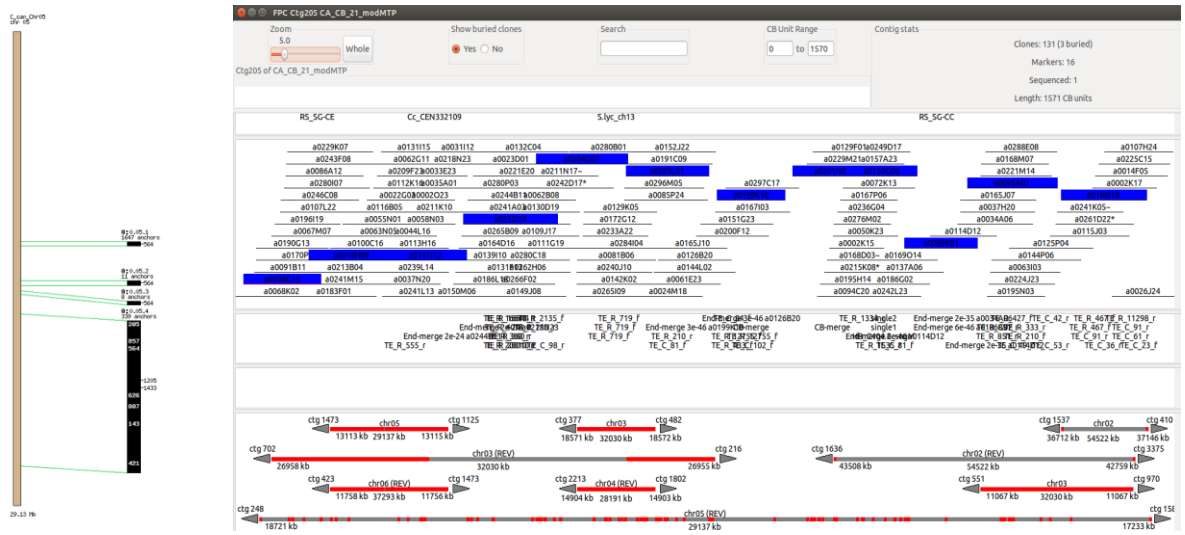


Figura 11. *Minimum Tiling Path*. A. Región Asociada a producción YLD2. B. MTP visualizado en el Mapa Físico.

El ensamblaje de los clones BAC para la región asociada al rasgo YLD2, obtuvo una secuencia de un tamaño de 7,38 Mb, que en *C. canephora representa* una región de 8,66 Mb (Figura 11). De 12 *contigs* provenientes del mapa físico calculados como constituyentes del MTP, se utilizaron 11 en el proceso de ensamblaje, ya que el CTG807-1 era sobrelapante con el CTG626-1 y correspondería, según la asignación de secuencias a otro subgenoma (E^a), ya que en este MTP el subgenoma predominante fue el C^a el cual fue el subgenoma ensamblado para la región (Tabla 5). Seis de los 11 *contigs* del FPC incluidos en el ensamblaje produjeron una secuencia continua, sin *gaps*, para los cinco restantes se tuvo que seguir la estrategia de *scaffolding* y les fueron agregados *gaps* de 100 Ns para obtener una sóla secuencia y que el proceso de ensamblaje de la región completa siguiera su curso. Al igual que en el caso de CTG807-1 se podría asignar a E^a, algunos de los BACs miembros de un *contig* FPC, razón por la cual algunos de ellos también fueron excluídos del ensamblaje al presentar un sobrelapamiento superior al 80% con el *contig* incluido en el ensamblaje.

Tabla 5. Ensamblaje BAC por BAC de la región YLD2 y sus estadísticas.

ORDEN FPC REGIÓN	CTGsFPC	No. Clones BACs	Secuencias Obtenidas	Gaps por CTG FPC	Longitud Inicial (Mb)	Longitud Final (Mb)
1	CTG205-0	18	5	4	2.82	1.41
2	CTG857-1	8	1	0	1.27	0.67
3	CTG14-1	11	8	7	6.88	1.93
4	CTG564-1	3	1	0	0.52	0.24
5	CTG1205-1	4	1	0	0.61	0.43
6	CTG379-1	10	2	1	1.56	1.13
7	CTG1433-1	8	1	0	1.25	0.79
8	CTG626-1	7	1	0	0.99	0.66
9	CTG807-1*	16	ND	ND	2.13	ND
10	CTG143-1	42	8	7	8.81	4.15
11	CTG2114-1	2	1	0	0.37	0.25

12	CTG421-1	13	4	3	2.39	1.28

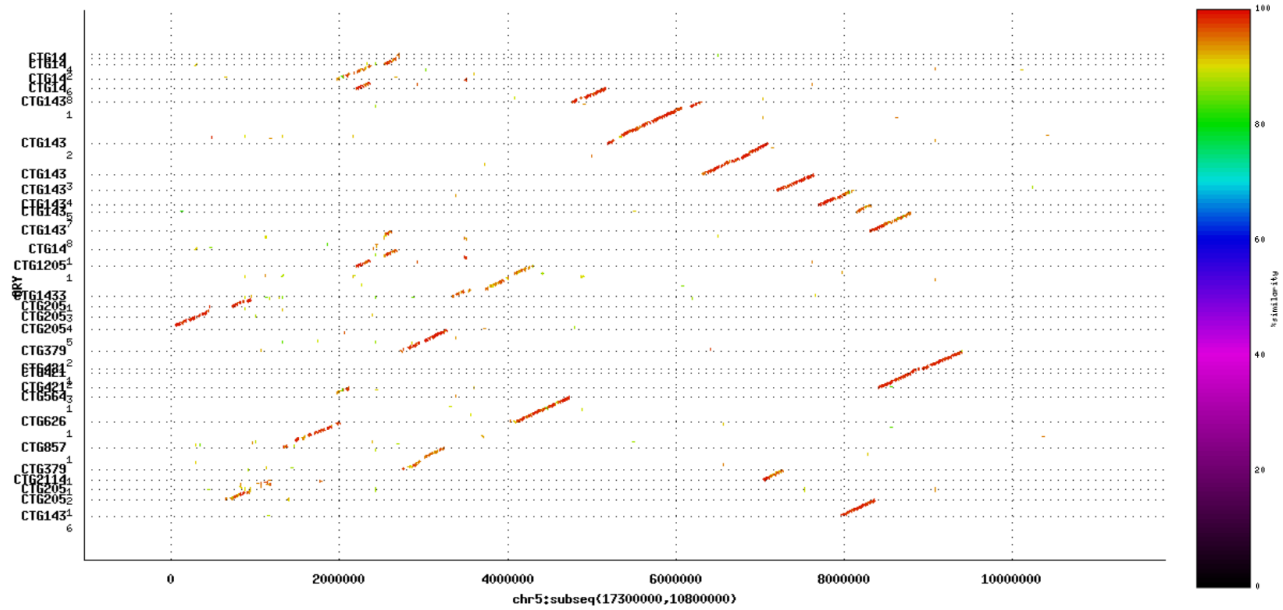


Figura 12. Dot plot de Región YLD2 en *C. canephora* contra los Contigs FPC pertenecientes al MTP.

La cobertura de la región YLD2 fue completa respecto al genoma de referencia (Figura 12), sin embargo, al hacer el ensamblaje y compararlo con esta región en el genoma de referencia, se pudo observar que en el *dot plot* de los ensamblajes de las secuencias de los subgenomas y la secuencia BAC existe un *indel* (Figura 13-A círculo rojo). También puede observarse una duplicación al final del ensamblaje donde se observan dos secuencias ocupando la misma región (Figura 13-B círculo rojo).

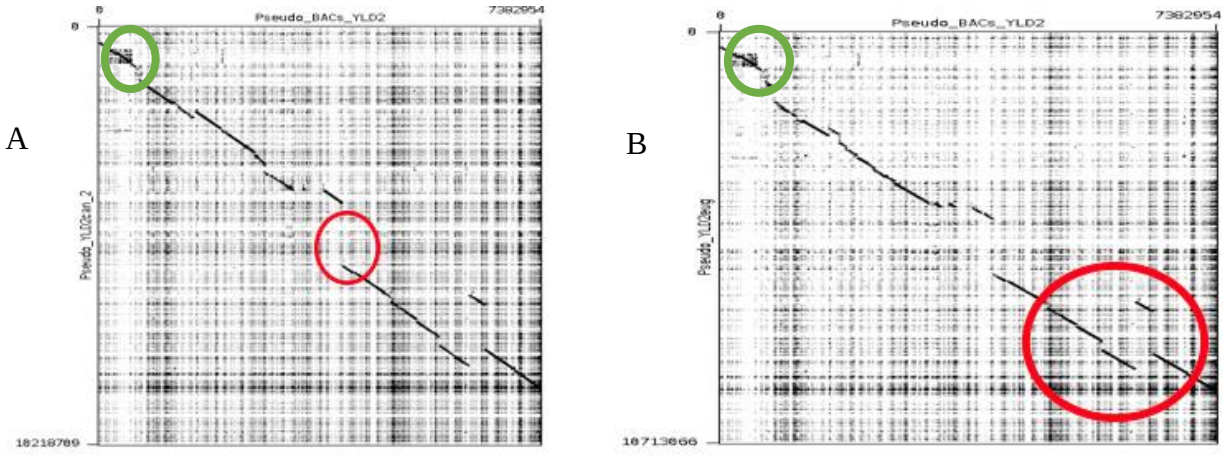


Figura 13. Dot plot de las comparaciones entre los scaffolds de los subgenomas y el ensamblaje BAC por BAC de la región YLD2. A. Subgenoma C[a]. B. Subgenoma E[a].

Al inicio de la secuencia en el ensamblaje WGS de los subgenomas y de los BACs de la región YLD2 presentan un resultado típico de una región rica en repeticiones que han sido consistentes en todos los *dot plots* de los subgenomas ensamblados a partir de WGS, así como los ensamblajes BAC por BAC, y el mismo ensamblaje contra el genoma de referencia (Figura 13-A y B círculo verde).

En el ensamblaje de la región PH5, se obtuvo un ensamblaje final de 10,18 Mb para el ensamblaje C^a y 11,89 Mb para E^a. La secuencia homologa *C. canephora* tiene una longitud de 9,88 Mb en (Figura 14). De 13 *contigs* provenientes del mapa físico calculados como constituyentes del MTP, se utilizaron 11 en el proceso de ensamblaje, ya que el CTG1616 y el CTG1139 no fueron incluidos en el mismo por dos razones. La primera es que al tratar de realizar el ensamblaje con el software minimus2 no existió un sobrelapamiento entre *contigs* FPC y por lo tanto no se pudo incluir en la secuencia final. La segunda, fue debido a que el resultado del alineamiento con *C. canephora* no fue significativo lo cual no permite ubicarla dentro de la región, Sin embargo, este resultado podría tratarse de una diferencia interespecífica por lo cual debe buscarse más evidencias, y obtener tal vez la región sobrelapante que sería prueba de la pertenencia de estos dos contigs con la región. De otro lado, como se puede observar en la figura 14, varios de los *contigs* FPC correspondían a los subgenomas C^a y E^a con sobrelapamientos del 100% de algunos de los *contigs*. Por tal motivo, se realizó el ensamblaje de esta región para los dos subgenomas. Los *contigs* FPC que fueron incluidos en el ensamblaje de PH5 C^a fueron CTG4461, CTG491, CTG855, CTG1341, CTG584, CTG894 y CTG743. Los *contig* CTG491, CTG1580, CTG588, CTG1203, CTG1538 y CTG743 hicieron parte del ensamblaje de esta región para el subgenoma E^a (Tabla 6). Nueve de los 11 *contigs* del FPC incluidos en el ensamblaje arrojaron como resultado del ensamblaje de los BACs una sola secuencia, para los tres restantes, cuyo resultado fue más de una secuencia después del proceso de ensamblaje, se asignaron a sus respectivos subgenomas y fueron ensamblados de acuerdo con esta característica. Finalmente se siguió la estrategia de *scaffolding* y les fueron agregados las 100 Ns para obtener una sóla secuencia y que el proceso de ensamblaje de la región completa siguiera su curso.

Tabla 6. Ensamblaje BAC por BAC de la región PH5 y sus estadísticas.

ORDEN FPC REGIÓN	CTGs FPC	NO. CLONES BACs	SECUENCIAS OBTENIDAS	Gaps POR CTG FPC	LONGITUD INICIAL (Mb)	LONGITUD FINAL (Mb)
1	CTG4461*	2	1	0	3.66	0.34
2	CTG491*	16	6	5	3.56	2.35
3	CTG855*	6	1	0	1.63	0.66
4	CTG1580	6	1	0	0.79	0.46
5	CTG1341*	8	1	0	1.10	0.65
6	CTG588	11	4	3	1.54	1.24
7	CTG584*	16	1	0	2.54	1.69
8	CTG1203	7	1	0	1.84	0.80
9	CTG894*	5	1	0	1.52	0.53
10	CTG1538	9	1	0	2.47	0.94
11	CTG1616**	6	1	0	1.67	0.58

12	CTG743	16	4	3	2.29	1.43
13	CTG1139**	8	2	1	1.21	0.81

* *Contigs* que corresponden al ensanblaje de la región PH5 del subgenoma C^a; ** Contigs excluidos del ensamblaje; los demás contigs fueron ensamblados junto a CTG491 para el subgenoma E^a.

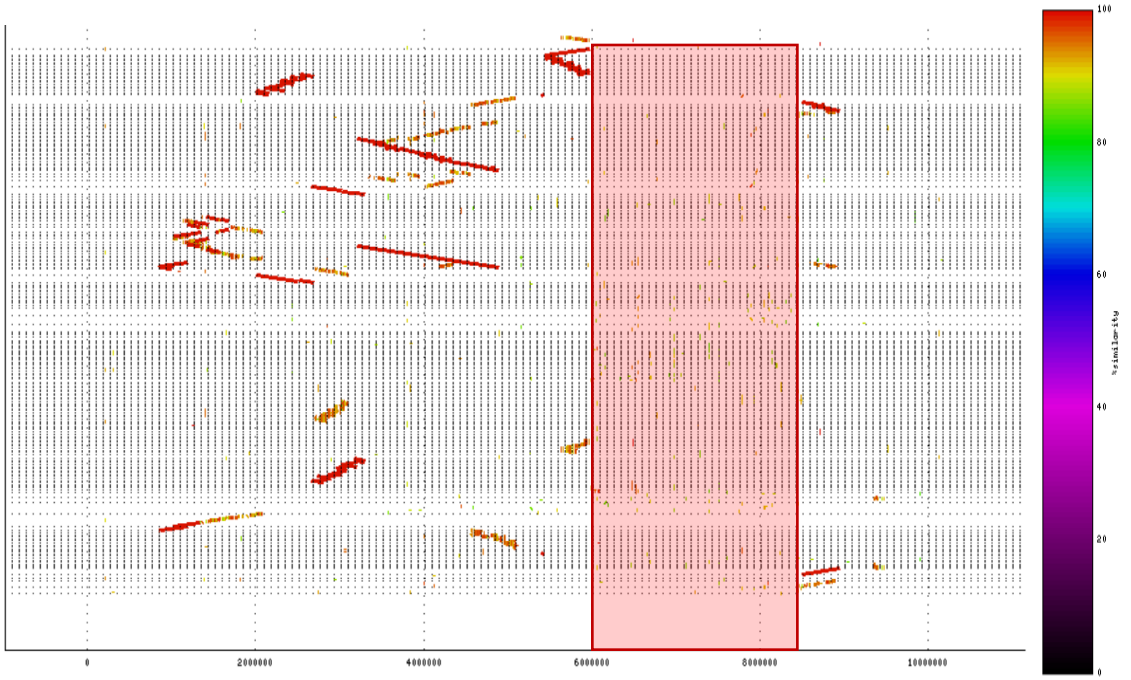


Figura 14. Dot plot de Región PH5 en *C. canephora* contra los Contigs FPC pertenecientes al MTP.

La cobertura de la región no fue completa (Figura 14), puede observarse como en las coordenadas de ~6.000.000 pb – 8.000.000 pb y de ~9.000.000 hasta el final de las secuencia correspondiente a la región en el genoma de referencia correspondientes a los *contigs* excluidos del ensamblaje por una similitud muy baja y un alineamiento fragmentado (Figura 14 círculo rojo) lo cual se puede observar en la representación del alineamiento en el *dot plot*, donde se ven líneas muy cortas de color verde y amarillo es de entre el 80 al 90% de acuerdo al código de color que se observa a la derecha en la figura 14. Al realizar las comparaciones de las regiones ensambladas para los subgenomas y su contraparte en *C. canephora* se observan tres *indels* cuando se comparan el subgenoma C^a y la referencia (Figura 15-A círculos rojos), mientras en el subgenoma E^a se observan la misma cantidad de indels, pero uno de ellos ubicado al inicio del ensamblaje obtenido por la estrategia BAC por BAC (Figura 15-B cuadros en rojo).

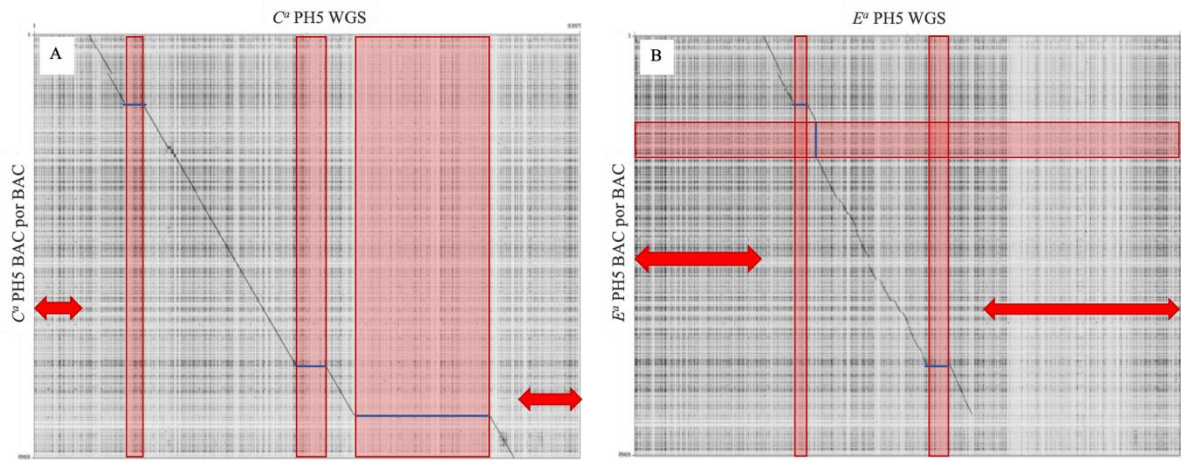


Figura 15. *Dotplot* de las comparaciones entre los *scaffolds* de los subgenomas WGS (Eje Horizontal) y el ensamblaje BAC por BAC (Eje Vertical) de la región PH5 A. Subgenoma C[a]. B. Subgenoma E[a].

Si se comparan los métodos probados en este trabajo respecto al ensamblaje de regiones de interés agronómico de una especie poliploide como el café, la mejor selección sería el uso de la secuenciación WGS con tecnología TGS, en este caso PacBio. De los dos fragmentos que fueron secuenciados con la metodología BAC por BAC, y a la ausencia de quimeras en los ensamblajes debido a, muy probablemente, la presencia de más de un haplotipo en el MTP. Esto podría explicarse debido a que el café tiene un genoma tetraploide y de acuerdo con la tecnología utilizada (FPC) es casi imposible realizar la separación de los haplotipos *a priori* para tener un buen ensamblaje con ausencia de quimerismo, entendido como presencia de secuencias de los dos subgenomas en el genoma de referencia. Esto a pesar de que los clones que pertenecen al MTP comparten tres bandas del *fingerprint*, lo que implica que las secuencias de los dos subgenomas son muy similares. Sin embargo, al comparar estas regiones con sus homólogas en *C. canephora* se puede observar una mayor similitud con algunos contigs lo que se infiere es el equivalente al subgenoma C[a]. Esto puede observarse en la figura 14 donde se ven dos secuencias mapeadas a la misma región de la referencia, pero con distintos porcentajes de similitud unas más cercana al 100% de similitud (presuntamente C[a]) y otras más cercanas al 90% en promedio (presuntamente E[a]).

Para evitar estos problemas se debe entender que existen diversas estrategias para la secuenciación y ensamblaje de genomas complejos como el café. Según Michael & VanBuren, (2015) en su artículo de revisión, en el ensamblaje de este tipo de genomas se aplican principalmente cuatro estrategias. La primera es la estrategia WGS el cual produce un ensamblaje muy fragmentado, ya que existen lecturas que incluyen más de una vez la misma región con las variantes caracteristiscas de regiones heterocigóticas, con haplotipos, y a la alta presencia de regiones repetitivas muy común en este tipo de genomas, donde en C. canephora presenta el 50% de este tipo de secuencias (Denoeud *et al*., 2014). Estas repeticiones crean "burbujas" en el gráfico de ensamblaje, interrumpiendo los grafos generados. Una segunda estrategia es la de usar un haploide doble, el cual es la mejor de las estrategias para obtener un genoma simplificado ya que sólo lee un haplotipo y el ensamblaje sólo debe lidiar con las secuencias repetitivas. El tercero, al igual que en este estudio, es el

uso de TGS, como PacBio, para ensamblar los haplotipos por separado, obteniendo secuencias de subgéneros completos. Sin embargo, deben usarse lecturas cortas de Illumina junto a PacBio para mejorar el ensamblaje. El cuarto es el típico enfoque BAC por BAC donde se secuencia un MTP, el cual sin embargo será una representación quimérica de los subgenomas presentes en el poliploide a pesar de la reducción de la complejidad, aún contendrán un alto porcentaje de repeticiones. Desde el punto de vista evolutivo, en café, este género presenta un punto de radiación tan sólo hace 0,5 millones de años, donde *Coffea arabica* tiene un punto de especiación calculada de ~0,1 millones de años, donde la especiación del género Coffea se dio en el pleistoceno (2Myr – 0.01 Myr) (Crops2008, n.d.). Por esta razón, no se esperaría que la divergencia de los subgenomas presentes en el alotetraploide *C. arabica* sea muy baja, haciendo que el essamblaje por este último método sea difícil de establecer de manera eficiente, ya que el riesgo de incluir en la misma secuencia segmentos correspondientes a los dos subgenomas es alto.

Además de los problemas de quimerismo, ya ha sido documentado por Videndi *et al.*, (2016), al igual que otros autores, que el ensamblaje BAC por BAC es uno de las estrategias de secuenciación más exitosas, ya que generalmente producen ensamblajes de mayor calidad que los WGS hasta la aparición de las tecnologías TGS, pero el costo de creación de las bibliotecas BACs, la generación de los perfiles de restricción de los clones BACs y la secuenciación de grandes cantidades de BACs del MTP hacen que los costos de secuenciación utilizando esta metodología sean muy altos. Por esta razón la disminución de costos de la secuenciación BAC por BAC es una prioridad entre muchos investigadores, buscando alternativas para el ensamblaje de genomas poliploides de alta calidad, permitiendo la aplicación de este enfoque para diferentes especies. Con este objetivo, Visendi *et al.*, (2016), implementó una estrategia para secuenciar y ensamblar el genoma de caña de azúcar y el brazo 7DS de trigo. Esto se logró optimizando la generación y el análisis bioinformático de los datos, mediante el uso de la secuenciación secuenciación Illumina 150 PE con bibliotecas MP con insertos de 6 a 10 kb con una cobertura de 100X por BAC de 96 grupos de 4 BACs a una profundidad de 400X. También fueron generados fragmento BES mediante el uso de secuenciación SANGER. En el artículo se aclara que "*En comparación con la secuenciación shotgun, una limitación generalmente reportada de los enfoques BAC-by-BAC es el costo. Sin embargo, esta comparación ignora los costos adicionales incurridos durante los proyectos de clonación de genes debido a conjuntos de genoma completo incompletos y altamente fragmentados y el costo de anclar y mejorar los asamblajes. Hemos demostrado a través de la indexación de BAC agrupados que los costos se pueden reducir sustancialmente al generar conjuntos de alta calidad. La cobertura física es fundamental para un montaje preciso. Hemos demostrado que los pools con cobertura física desigual se pueden identificar, reensamblar y volver a ensamblar fácilmente con una cobertura suficiente*." (Visendi *et al*., 2016) p 5-6.

Según Bolger *et al.*, (2014) y Visendi *et al.*, (2016), la secuenciación WGS, para genomas grandes, repetitivos y complejos dan como resultado ensamblajes fragmentados que requieren esfuerzos adicionales para ordenar, orientar y producir pseudomoléculas aceptables. Estos datos además deben incluir datos MP y de lectura larga, así como mapas físicos y mapas genéticos, de especies estrechamente relacionadas con genomas de referencia, genotipificación por secuenciación y secuenciación poblacional. Sin embargo, Michael *et al.*, (2018), obtuvieron un genoma ensamblado de alta contigüidad de *A. thaliana* usando el secuenciador TGS MinION, donde pudieron obtener pseudomoléculas correspondientes a brazos del cromosoma, 62 *contigs* con un N50 de longitud de 12,3 Mb.

Esta secuenciación permitió la resolución de un QTL, que había sido recalcitrante para la aproximación BAC por BAC basada en SANGER, demostrando que el WGS se está convirtiendo en el camino más simple y barato para comprender la variación estructural de una región sencilla del genoma, el cual es uno de los principales conclusiones de ese trabajo y de este, en el cual también se utilizó una metodología TGS (PacBio) (Michael *et al*., 2018).

En este trabajo se pudo observar que una estrategia WGS exitosa debe incluir TGS, ya sea secuenciación PacBio, u Oxford Nanopore que se podría usar como alternativa. Como se pudo observar se pueden clasificar las secuencias de lecturas largas en subgenomas, lo cual mejora el hecho de la aparición de quimeras dada la presencia de los dos subgenomas en una sola secuencia y la longitud de las secuencias da una alta contigüidad. Para mejorar aún más la calidad del genoma obtenido se puede usar secuenciación Illumina para corregir errores y usando métodos de scaffolding como el mapeo óptico.

4.3. Anotación de los genes contenidos en QTLs de importancia agronómica en café.

Se predijeron los genes correspondientes a la secuencia ensamblada WGS para los subgenoma C^a y E^a de las regiones de importancia agronómica con *maker*. Se encontraron más genes en promedio para las secuencias pertenecientes a los subgenomas E^a que el del subgenoma C^a. De estos genes anotados, lo más representativos fueron secuencias asociadas con resistencia y defensa, como genes con los dominos NB-ARC, p-loop, regiones ricas en leucina y monoxigenasas P450. Algunos de ellos a una distancia de 11.8 cM o alrededor de 1.9 Mb en la región YLD2, 0.59 cM ó ~620 kb en PH5, 2.92 cM ó 2Mb para PH1.

4.3.1. Anotación de genes codificantes

Cinco regiones de seis contenían los marcadores que tenían una correlación significativa con el rasgo. Estos fueron BS4, PH1, PH4, PH5 y YLD2. De estas regiones, fueron mapeados los marcadores asociados al rasgo en los dos subgenomas para PH1, PH5 y YLD2. Para las regiones BS4 y PH4 sólo hubo un *hit contra* E^a. En la tabla 8 se pueden observar los genes y cómo están distribuidos de acuerdo con la distancia física al marcador correlacionado al rasgo. Como no se mapearon todos los marcadores dentro de las regiones de interés, se seleccionaron aquellos que estuvieran dentro de un rango < a 30cM a lado y lado del marcador asociado a las diferentes regiones. Es así como para la región BS4 E^a dentro de la región flanqueada por marcadores encontramos 31 genes, PH1 C^a 93, PH1 E^a 118, PH4 E^a 71, PH5 C^a 140, PH5 C^a 149, YLD2 C^a 218 y YLD E^a 189 genes (tabla 7).

En el caso de los genes compartidos por la mayoría de las regiones, se encontraron los genes Pquinasas (Tabla 7). A pesar de tener un papel preponderante en muchos mecanismos de regulación celular, las quinasas son una superfamilia que representan alrededor de 1000 miembros en *A. thaliana*. Las funciones biológicas de las proteínas quinasas vegetales incluyen la resistencia, como en el caso de la proteína PTO, la cual tiene un dominio quinasa y otorga resistencia a el tomate un patógeno bacteriano, también está implicada en la regulación negativa de la señalización del etileno en la planta (CTR1- TKL, Tirosin-quinasa), la alteración en el desarrollo de los órganos florales (TLK - Quinasa Tousled). Entre otras funciones que se han descrito se encuentra que las proteínas quinasas vegetales son componentes de las redes de señalización, como en la percepción de los agentes bióticos, la calidad y cantidad de la luz, las hormonas vegetales y diversas condiciones

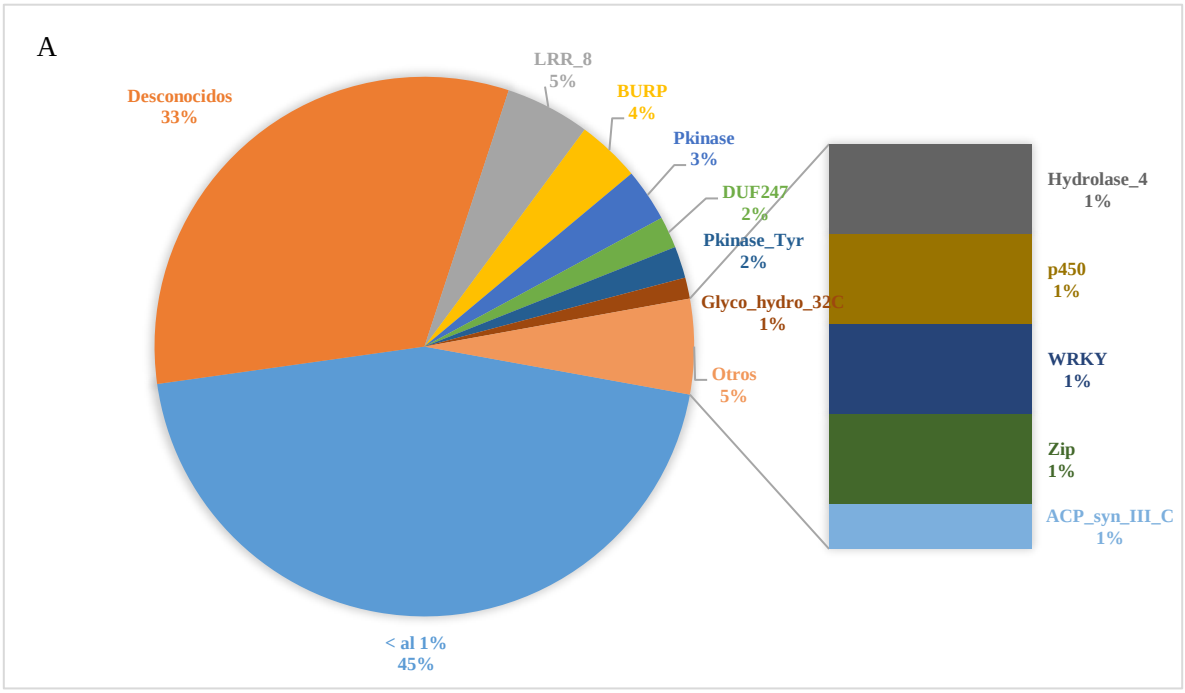
ambientales adversas. Además se han asociado con la regulación diurna y circadiana, la regulación del ciclo celular, los procesos de desarrollo, el transporte modulatorio de vesículas y las actividades del canal de proteínas, y en la regulación del metabolismo celular (Lehti-Shiu & Shiu, 2012).

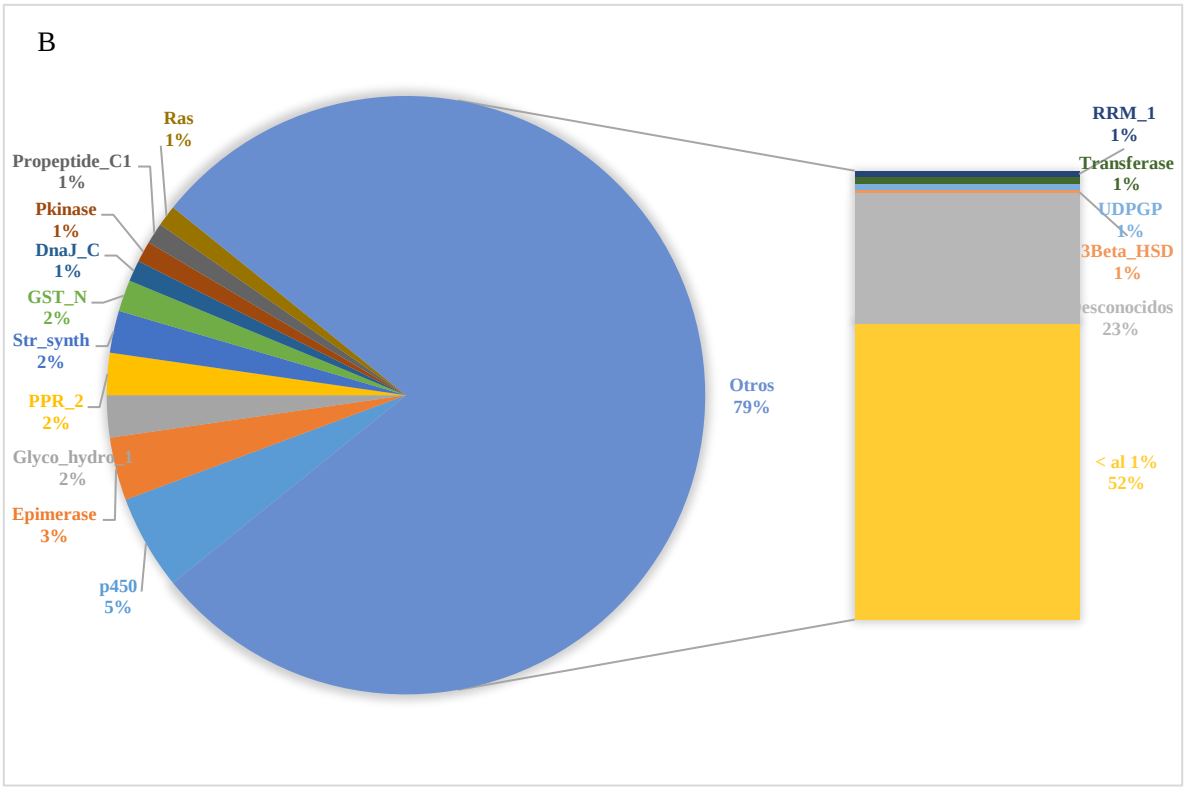
Dentro de los genes más representativos encontrados en todas las regiones, se encuentran genes asociados a resistencia y defensa como genes con los dominos NB-ARC ó su anotación alternativa NBS-LRR, p-loop, regiones ricas en leucina y monoxigenasas P450. Para las regiones altura de la planta, se encontraron familias de genes WD40, que pueden estar relacionados con la resistencia a estrés abiótico. Estos también desempeñan un papel crucial en diversas interacciones proteína-proteína, actuando como ensambladores moléculares y, por tanto, ayudando a la adecuada actividad de las proteínas. Un subconjunto de proteínas WD40 han sido nombrados como DWD [Unión de ADN dañado (DDB) WD40] con base en su interacción con DDB1 y CULLIN4 (CUL4). Las E3 ubiquitin CUL4-DDB1 ligasas usan proteínas DWD como adaptadores moleculares para el reconocimiento de sustratos, y modulan múltiples procesos biológicos a través de la proteólisis dependiente de ubiquitina como el mecanismo de reparación del ADN causado por daño UV y metilación de histonas (modificación postraduccional) (A. K. Mishra, Muthamilarasan, Khan, Parida, & Prasad, 2014).

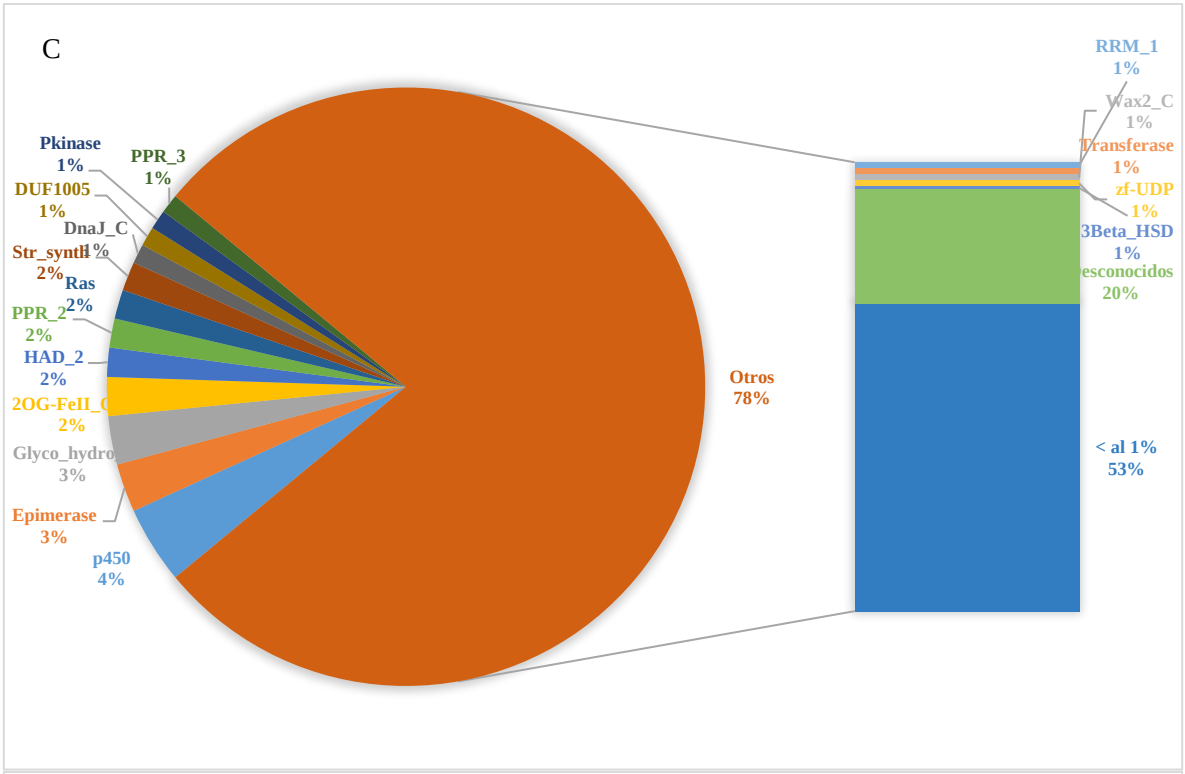
En las plantas, todas las regiones a excepción de BS4, se anotaron los pentatricopéptidos (PPR) y la superfamilia Ras (Tabla 7, anexo 3). Los primeros están relacionados con la reproducción, ya que los mutantes de las proteínas mitocondriales PPR a menudo presentan defectos en el desarrollo embrionario y la esterilidad masculina citoplasmática. Los genomas mitocondriales de plantas a menudo codifican factores que inducen la esterilidad masculina citoplasmática y la incapacidad para producir polen viable. El restaurador de las proteínas PPR de la fertilidad parece que funciona mediante la prevención de la traducción o acumulación de transcritos mitocondriales, cuyos productos génicos inducen esterilidad masculina citoplasmática. Los mutantes de PPR de plantas a menudo no tienen ningún mecanismo para reemplazar estos factores mitocondriales que inducen la esterilidad masculina citoplasmática, y por lo tanto no pueden producir polen funcional. Los mutantes de proteínas PPR de cloroplastos en las plantas también presentan varios fenotipos, incluyendo el desarrollo de embriones defectuosos, la capacidad fotosintética aberrante y los cambios en la pigmentación de la semilla, lo cual tienen implicaciones significativas en la agricultura (Manna, 2015). Estos genes tienen una directa significancia con la producción de semilla y crecimiento y formación de las paredes celulares y los frutos. Los genes con dominios ras en cambio se unen a reguladores de genes y consecuentemente a la señalización celular, los cuales incluyen varias familias como la Ras, Rab, Rho, Arf/Sar, y Ran. Los miembros de la familia Ras funcionan como nodos de señalización que se activan mediante diversos estímulos extracelulares y que regulan la señalización intracelular. Esta señalización controla en última instancia la transcripción génica, que a su vez influye en los procesos fundamentales, como el crecimiento y la diferenciación celular (Rojas, Fuentes, Rausell, & Valencia, 2012).

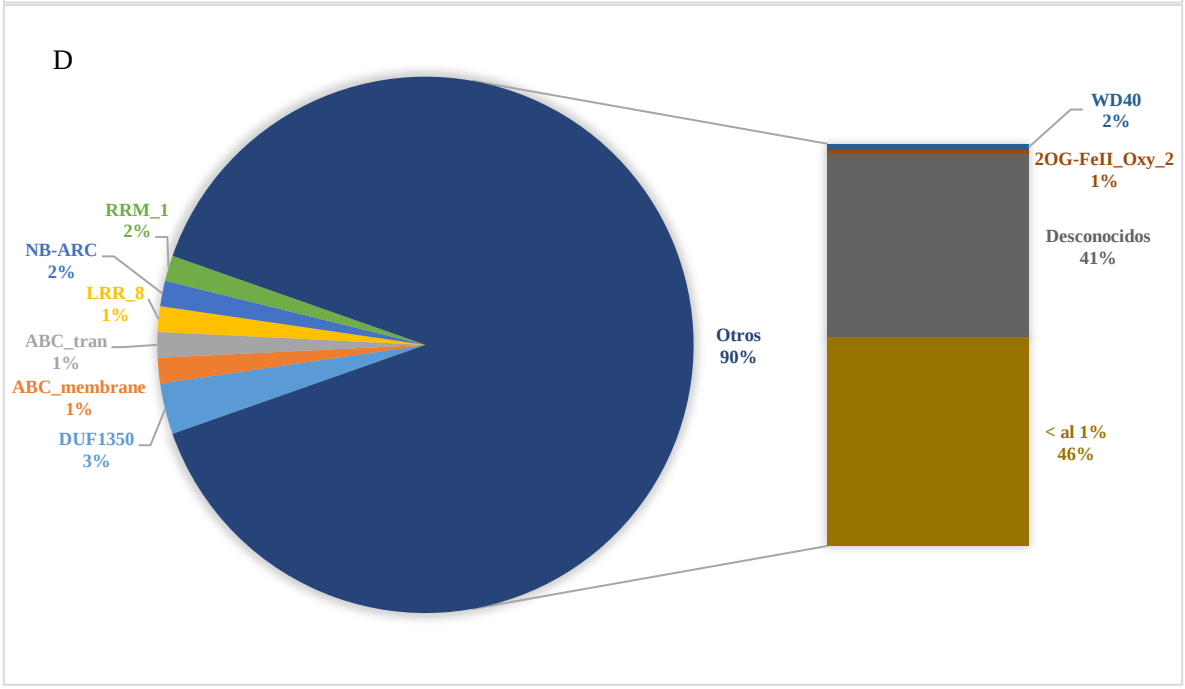
Para la región BS4 sólo se describieron 26 genes únicos, mientras PH1 presenta 11 genes para C^a y 27 para E^a, la región PH4 E^a 49, para la región PH5 C^a 33, E^a 41 y finalmente para la región YLD2 27 genes para C^a y 14 para E^a. La mayoría de los genes encontrados que son exclusivos para cada una de las regiones, están relacionados con factores de transcripción, involucrados en la síntesis de pared celular. Sin embargo, el porcentaje de secuencias aún sin anotar es del 33% en la región BS4, 23% en PH1 C^a y 20% en PH1 E^a, la

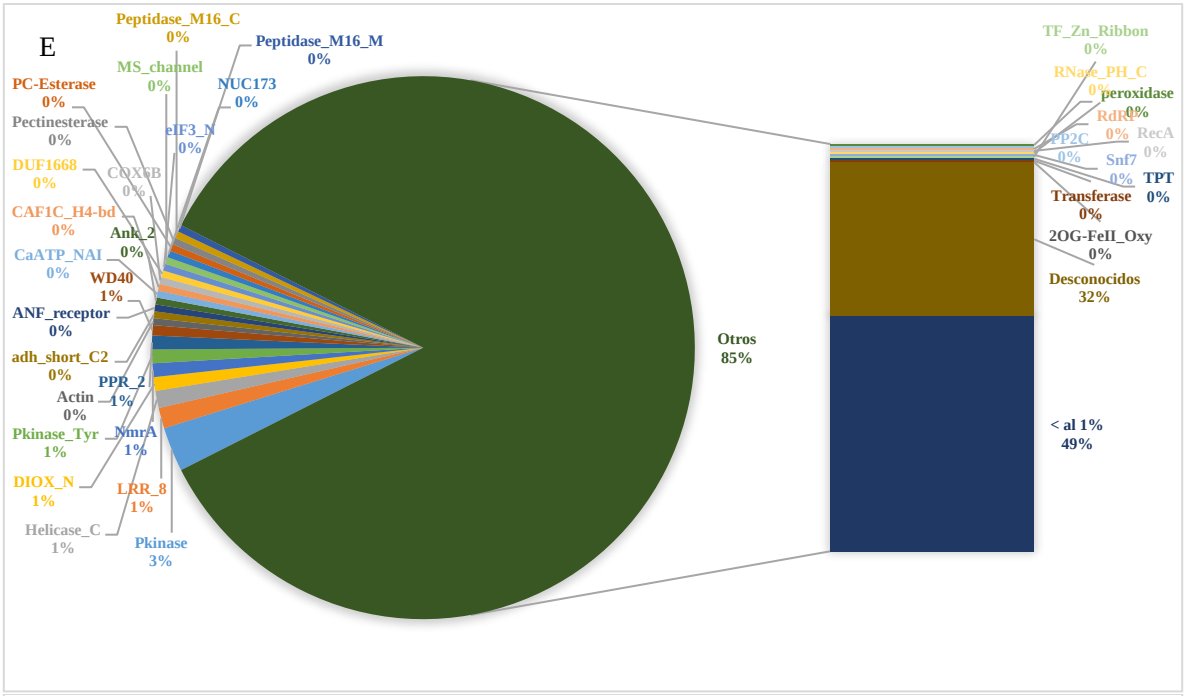
región PH4 Eª tiene el 41% de genes no anotados, la región PH5 Cª el 32 y Eª el 24%, y la región YLD2 22% para Cª y 41% para Eª (Figura 19).

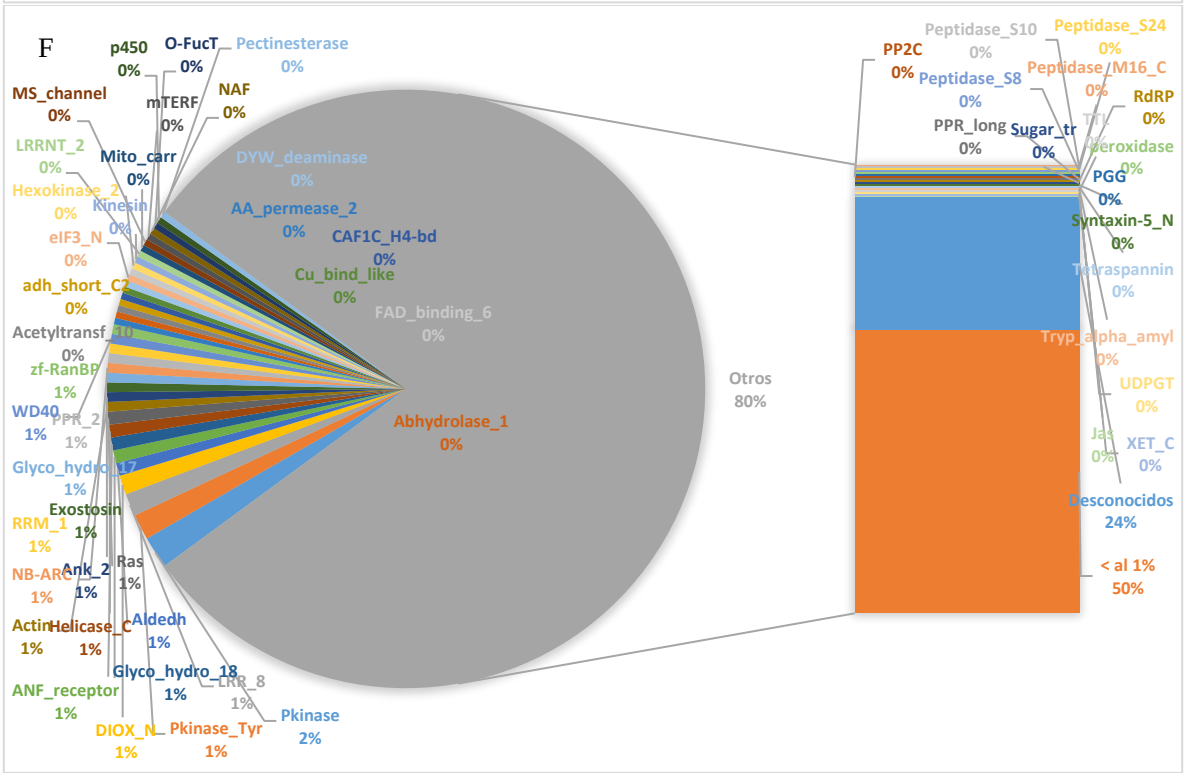





C
Pkinase 1%
PPR_3 1%
DUF1005 1%
DnaJ_C 1%
Str_synth 2%
Ras 2%
PPR_2 2%
HAD_2 2%
2OG-FeII_C 2%
Glyco_hydro 3%
Epimerase 3%
p450 4%
Otros 78%
RRM_1 1%
Wax2_C 1%
Transferase 1%
zf-UDP 1%
3Beta_HSD 1%
esconocidos 20%
< al 1% 53%


D
RRM_1 2%
NB-ARC 2%
LRR_8 1%
ABC_tran 1%
ABC_membrane 1%
DUF1350 3%
Otros 90%
WD40 2%
2OG-FeII_Oxy_2 1%
Desconocidos 41%
< al 1% 46%


E
Peptidase_M16_C 0%
Peptidase_M16_M 0%
MS_channel 0%
NUC173 0%
PC-Esterase 0%
Pectinesterase 0%
eIF3_N 0%
DUF1668 0%
COX6B 0%
CAF1C_H4-bd 0%
Ank_2 0%
CaATP_NAI 0%
WD40 1%
ANF_receptor 0%
adh_short_C2 0%
PPR_2 1%
Actin 0%
Pkinase_Tyr 1%
NmrA 1%
DIOX_N 1%
LRR_8 1%
Helicase_C 1%
Pkinase 3%
Otros 85%
TF_Zn_Ribbon 0%
RNase_PH_C 0%
peroxidase 0%
RdRP 0%
PP2C 0%
RecA 0%
Snf7 0%
TPT 0%
Transferase 0%
2OG-FeII_Oxy 0%
Desconocidos 32%
< al 1% 49%


F
p450 0%
O-FucT 0%
Pectinesterase 0%
MS_channel 0%
mTERF 0%
NAF 0%
LRRNT_2 0%
Mito_carr 0%
DYW_deaminase 0%
Hexokinase_2 0%
Kinesin 0%
AA_permease_2 0%
CAF1C_H4-bd 0%
eIF3_N 0%
Cu_bind_like 0%
adh_short_C2 0%
FAD_binding_6 0%
Acetyltransf_10 0%
zf-RanBP 1%
WD40 1%
PPR_2 1%
Abhydrolase_1 0%
Glyco_hydro_17 1%
Exostosin 1%
RRM_1 1%
NB-ARC 1%
Ank_2 1%
Ras 1%
Actin 1%
Helicase_C 1%
Aldedh 1%
Glyco_hydro_18 1%
LRR_8 1%
ANF_receptor 1%
Pkinase 2%
DIOX_N 1%
Pkinase_Tyr 1%
Otros 80%
Peptidase_S10 0%
Peptidase_S24 0%
PP2C 0%
Peptidase_M16_C 0%
Peptidase_S8 0%
RdRP 0%
PPR_long 0%
Sugar_tr 0%
TPT 0%
peroxidase 0%
PGG 0%
Syntaxin-5_N 0%
Tetraspannin 0%
Tryp_alpha_amyl 0%
UDPGT 0%
Jas 0%
XET_C 0%
Desconocidos 24%
< al 1% 50%

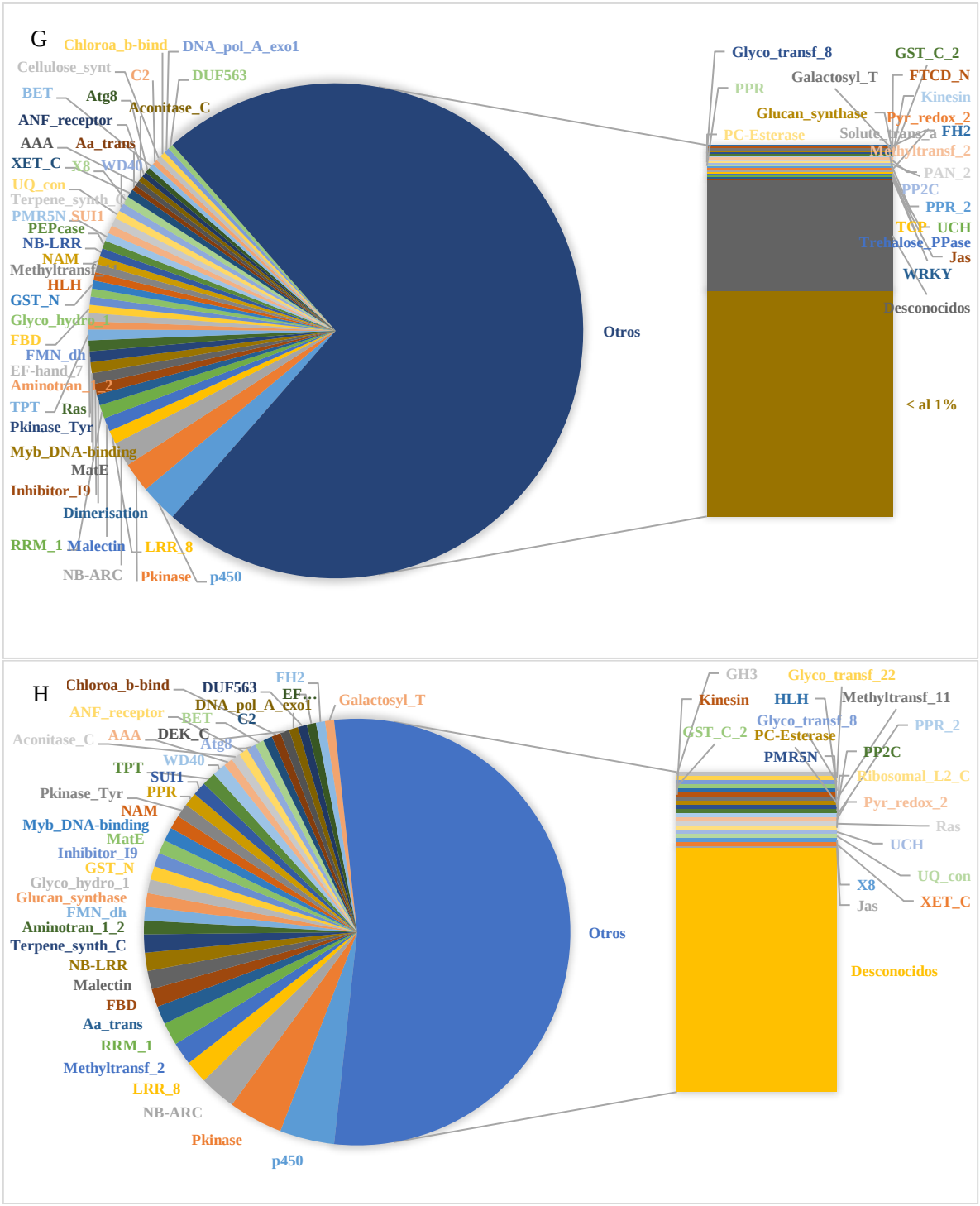


Figura 16. Anotación del PFAM. A. BS4 *C*[a]; B. PH1 *C*[a]; C. PH1 E[a]; D. PH4 E[a]; E. PH5 *C*[a]; F. PH5 E[a]; G. YLD2 *C*[a]; H. YLD2 E[a].

Tabla 7. Genes dentro de los rangos de marcadores moleculares mapeados por región

BS4 E[a]		PH1 C[a]		PH1 E[a]		PH4 E[a]		PH5 C[a]		PH5 E[a]		YLD2 C[a]		YLD2 E[a]	
Genes	**No. De Genes**	**Genes**	**No. De Genes**	**Genes**	**No. De Genes**	**Genes**	**No. De Genes**	**Genes**	**No. De Genes**	**Genes**	**No. De Genes**	**Genes**	**No. De Genes**	**Genes**	**No. De Genes**
Pkinase	4	p450	8	p450	7	DUF1350	4	Pkinase	8	Glyco_hydro_18	4	p450	14	p450	12
Pkinase_tyr	1	Glyco_hydro_1	4	Glyco_hydro_1	5	ABC_membrane	2	NmrA	4	LRR_8	3	NB-ARC	6	Pkinase	6
Uq_con	1	GST_N	3	HAD_2	3	LRR_8	2	LRR_8	3	NB-ARC	3	Pkinase	5	NB-ARC	5
Rpe65	1	PPR_2	2	Ras	3	RRM_1	2	Transferase	2	Pkinase_Tyr	3	Malectin	5	Methyltransf_2	5
Ech_1	1	DnaJ_C	2	Pkinase	2	NB-ARC	2	RNase_PH_C	2	adh_short_C2	2	LRR_8	4	LRR_8	4
Peptidase_s10	1	Propeptide_C1	2	PPR_3	2	ABC_tran	2	TF_Zn_Ribbon	2	Helicase_C	2	Inhibitor_I9	4	Terpene_synth_C	4
Enod93	1	Ras	2	PPR_2	2	Hydrolase_4	1	Actin	2	DIOX_N	2	Ras	4	Malectin	4
Profilin	1	RRM_1	2	RRM_1	2	Cellulose_synt	1	Helicase_C	2	DYW_deaminase	2	Dimerisation	4	Aa_trans	4
Fad_binding_4	1	Transferase	2	DUF1005	2	DUF5110	1	PPR_2	2	Actin	2	Myb_DNA-binding	4	GST_N	3
Tocopherol_cycl	1	Pkinase	2	Wax2_C	2	START	1	eIF3_N	2	Pkinase	2	TPT	3	Myb_DNA-binding	3
Far1	1	UDPGP	2	zf-UDP	2	LSM14	1	CaATP_NAI	2	p450	2	PMR5N	3	Inhibitor_I9	3
Yippee-mis18	1	Pkinase_Tyr	1	Transferase	2	RWD	1	Peptidase_M16_M	2	Tryp_alpha_amyl	2	Pkinase_Tyr	3	FMN_dh	3
Gat	1	Sigma70_r2	1	RWD	1	Glyco_hydro_19	1	adh_short_C2	2	Abhydrolase_1	2	GST_N	3	Glucan_synthase	3
Duf810	1	Cys_Met_Meta_PP	1	CK_II_beta	1	RST	1	EMP70	1	eIF3_N	2	Terpene_synth_C	3	MatE	3
Pfk	1	ECH_2	1	NB-ARC	1	DIM1	1	LRR_6	1	Glyco_hydro_17	2	FMN_dh	3	PPR_2	2
Gata	1	OHCU_decarbox	1	DnaJ_C	1	RSN1_TM	1	Cytokin-bind	1	AA_permease_2	2	MatE	3	UCH	2
Pnp_udp_1	1	ELMO_CED12	1	PTR2	1	AOX	1	TPT	1	PCRF	1	XET_C	3	Ras	2
Hydrolase_3	1	Cu_bind_like	1	AATF-Che1	1	XET_C	1	Gpi1	1	ubiquitin	1	PPR_2	2	GST_C_2	2
Rcc1	1	ENT	1	ATP-synt_DE_N	1	JmjN	1	Tmemb_14	1	Reticulon	1	Aminotran_1_2	2	PPR	2
Hydrolase_4	1	AMP-binding_C	1	DUF498	1	RPE65	1	DUF3336	1	Choline_transpo	1	FTCD_N	2	ANF_receptor	2
Sawadee	1	FKS1_dom1	1	O-FucT	1	GPAT_N	1	zf-CCCH	1	Cation_efflux	1	Trehalose_PPase	2	Pyr_redox_2	2

Lrr_6	1	tRNA_synt_2f	1	DUF632	1	RNA_pol_Rpc34	1	Tryp_alpha_amyl	1	COX6B	1	GST_C_2	2	PC-Esterase	2				
Trna-synt_1b	1	G-patch	1	Cnd1_N	1	EamA	1	SUZ	1	Carb_anhydrase	1	UCH	2	TPT	2				
Malic_m	1	CK_II_beta	1	DYW_deaminase	1	RNA_pol_Rpb2_5	1	HSF_DNA-bind	1	Cytokin-bind	1	Pyr_redox_2	2	Pkinase_Tyr	2				
Yif1	1	ATP-synt_DE_N	1	Ribosomal_L3	1	ThiF	1	SRP-alpha_N	1	Spc97_Spc98	1	X8	2	XET_C	2				
Mekhla	1	Plus-3	1	ECH_2	1	Ribosomal_L35p	1	FA_desaturase	1	DAO	1	HLH	2	PMR5N	2				
Aminotran_4	1	Glyco_hydro_20	1	Stm1_N	1	COG6	1	Spc97_Spc98	1	Zein-binding	1	PEPcase	2	Aminotran_1_2	2				
Nucleopor_nup85	1	Cyclin	1	eIF2A	1	Ribosomal_L13e	1	DYW_deaminase	1	Dimerisation	1	Methyltransf_2	2	PP2C	1				
		Glyco_transf_22	1	MORN	1	BAH	1	Sina	1	OTU	1	ANF_receptor	2	X8	1				
		Ribosomal_L37ae	1	ELMO_CED12	1	PTEN_C2	1	DIOX_N	1	ABC_tran	1	UQ_con	2	Sof1	1				
		Glycohydro_20b2	1	Neprosin_AP	1	Syntaxin_2	1	RVT_3	1	Peptidase_S28	1	Aa_trans	2	DUF572	1				
		SCAI	1	ENT	1	PPTA	1	CH	1	DNA_pol3_delta2	1	Glucan_synthase	2	Peptidase_C13	1				
		Citrate_synt	1	CBFB_NFYA	1	Transglut_core2	1	zf-C6H2	1	PRT_C	1	PC-Esterase	2	DUF588	1				
		TP_methylase	1	F-box-like	1	Pkinase_Tyr	1	APO_RNA-bind	1	DnaJ	1	Methyltransf_11	2	Ribosomal_L13	1				
		HAD_2	1	Plant_NMP1	1	Iso_dh	1	Ribosomal_L21e	1	SAP	1	polyprenyl_synt	1	DUF604	1				
		UbiA	1	Fer4_7	1	PIG-X	1	LEA_1	1	DUF2048	1	Transketolase_C	1	Transketolase_C	1				
		HALZ	1	Pribosyl_synth	1	GWT1	1	RecA	1	Transferase	1	SBP	1	DUF641	1				
		OMPdecase	1	FKS1_dom1	1	PHM7_cyt	1	UCH	1	DUF3336	1	EF_TS	1	Ank_2	1				
		ILVD_EDD	1	RcbX	1	Glyco_hydro_31	1	QWRF	1	WRC	1	ANAPC4_WD40	1	DUF92	1				
		Cnd1_N	1	FMN_red	1	Peptidase_M49	1	FH2	1	DUF3474	1	Elongin_A	1	CS	1				
		LEA_1	1	Ribosomal_S7e	1	Exo_endo_phos	1	Pyr_redox_2	1	zf-C6H2	1	DUF455	1	ECH_2	1				
		Plant_NMP1	1	Fra10Ac1	1	PAP_fibrillin	1	Esterase	1	DUF4187	1	EMP70	1	Pribosyltran_N	1				
		LIN37	1	Sec23_BS	1	DUF647	1	PUNUT	1	Nudix_N_2	1	TB2_DP1_HVA22	1	EF_TS	1				
		PP-binding	1	GED	1	PAE	1	eIF-5a	1	DUF773	1	Epimerase	1	SBF	1				
		LSM	1	DDE_Tnp_4	1	DUF1664	1	PRP4	1	PC-Esterase	1	WSD	1	Elongin_A	1				
		PPR_3	1	ABC_membrane	1	Nucleotid_trans	1	DUF3479	1	Adenine_glyco	1	ERO1	1	ThylakoidFormat	1				
		MA3	1	Mob_synth_C	1	DUF1336	1	Zein-binding	1	Peptidase_M17_N	1	Amidase	1	EMP70	1				
		PTR2	1	Glyco_hydro_20	1	WD40	1	DNA_pol3_delta2	1	EF-hand_5	1	F-box	1	DUF1421	1				

		Metallopho s	1	N_Asn_am idohyd	1	DAD	1	PP2C	1	peroxidase	1	PPR_3	1	Epimerase	1				
		Ribosomal_ L3	1	Glyco_tran sf_22	1	NAD_bindi ng_10	1	DBINO	1	EF-hand_7	1	FBA_1	1	DUF3403	1				
		Mito_carr	1	NDK	1	CobW_C	1	Pkinase_Tyr	1	PPR_2	1	Ribosomal_ L5_C	1	ERO1	1				
		Ribosomal_ S7e	1	Glycohydr o_20b2	1	Myb_DNA- binding	1	Coq4	1	eIF-5a	1	FBA_3	1	PAM2	1				
		Mob_synth _C	1	NOG1	1	C1_2	1	Yippee- Mis18	1	Ras	1	Sof1	1	F-box	1				
		RWD	1	GST_N	1	ACBP	1	Carb_anhydr ase	1	Acetyltransf _10	1	Fer4_13	1	PGP_phosp hatase	1				
		Ubie_methy ltran	1	OMPdecas e	1	B3	1	peroxidase	1	RNase_PH_ C	1	TPR_1	1	FBA_1	1				
		Sec23_BS	1	ADH_zinc _N	1	Trehalose_P Pase	1	BRF1	1	EMP70	1	FH2	1	Cyclin_C	1				
		Ufm1	1	Palm_thioe st	1	Adaptin_N	1	Peptidase_S 28	1	SHMT	1	DUF641	1	FBD	1				
		Stm1_N	1	HALZ	1	zf-CCHC_6	1	Aldedh	1	ENT	1	FKS1_dom 1	1	Dimerisatio n	1				
		zf-CCCH_2	1	Pkinase_T yr	1	Med26	1	Peptidase_M 17_N	1	Tim17	1	zf-C3HC	1	FH2	1				
		DUF1005	1	HD-ZIP_N	1	Myb_DNA- bind_3	1	Longin	1	Esterase	1	ATP- synt_D	1	DnaJ	1				
		zf-LSD1	1	PP-binding	1	2OG- FeII_Oxy_2	1	YbaB_DNA _bd	1	Trigger_N	1	PfkB	1	FKS1_dom 1	1				
		Tryp_alpha _amyl	1	HLH	1	Myb_CC_L HEQLE	1	Ldh_1_N	1	Exostosin	1	Band_7	1	Ribosomal_ L5_C	1				
		Alg6_Alg8	1	Cpn10	1	MFMR_ass oc	1	PCRF	1	Utp11	1	Plug_transl ocon	1	BET	1				
		DYW_dea minase	1	ILVD_ED D	1			HLH	1	FA_desatur ase	1	Gar1	1	SelR	1				
		O-FucT	1	Propeptide _C1	1			PC-Esterase	1	YbaB_DNA _bd	1	PPR	1	FTCD_N	1				
		MORN	1	KH_1	1			HD-ZIP_N	1	FAD_bindi ng_3	1	BES1_N	1	SUI1	1				
		Vps62	1	Cu_bind_li ke	1			Nudix_N_2	1	zf-AN1	1	Pterin_4a	1	GH3	1				
		N_Asn_ami dohyd	1	LCM	1			Glyco_hydr o_17	1	FAD_bindi ng_6	1	Glyco_hydr o_14	1	DUF1298	1				
		zf-CCCH_3	1	Ribosomal _L18p	1			NUC173	1	zf-NOSIP	1	Ribosomal_ L1	1	Bystin	1				
		NB-ARC	1	Telomeras e_RBD	1			FAD_bindin g_3	1	Fip1	1	Glyco_hydr o_16	1	tRNA- synt_1b	1				
		zf-met	1	Ribosomal _L37ae	1			Nt_Gln_ami dase	1	Nse4_C	1	RRM_2	1	Glyco_hydr o_14	1				
		NDK	1	tRNA_synt _2f	1			Exo_endo_p hos	1	FPL	1	Glyco_trans f_20	1	Vps16_C	1				
		Neprosin_A P	1	Cyclin	1			NST1	1	O-FucT	1	SelR	1	Glyco_hydr o_16	1				
				Tryp_alpha _amyl	1			ENT	1	GDPD	1	Glyco_trans f_8	1	zf-C3HC	1				
				SCAI	1			WW	1	CAF1	1	SOUL	1	Glyco_trans f_8	1				

			Ubie_methyltran	1		ubiquitin	1	AHSA1	1	BET	1	NTP_transferase	1
			Sigma70_r4	1		NIF	1	PCO_ADO	1	DUF572	1	Chloroa_b-bind	1
			UPF0565	1		EF-hand_6	1	Aldedh	1	Bystin	1	PA	1
			Str_synth	1		Nic96	1	PDCD7	1	TPX2_importin	1	bZIP_2	1
			Alg6_Alg8	1		DUF563	1	Glycos_transf_4	1	GTP_EFTU_D2	1	CorA	1
			zf-CCCH_2	1		NEMP	1	Peptidase_S10	1	tRNA-synt_1b	1	GTP_EFTU_D2	1
			LEA_1	1		DUF3474	1	Gpi1	1	GUB_WAK_bind	1	PfkB	1
			tRNA-synt_2d	1		NB-ARC	1	Peptidase_S41	1	Vps16_C	1	GUB_WAK_bind	1
			LIN37	1		DnaJ	1	HD-ZIP_N	1	HAD_2	1	ATP-synt_D	1
			UbiA	1		NAF	1	CaM_binding	1	Dynamin_M	1	HATPase_c	1
			LSM	1		DNA_pol_E_B	1	Alg14	1	HATPase_c	1	Plug_translocon	1
			Ufm1	1		NAD_Gly3P_dh_C	1	PPR_1	1	zf-MYND	1	HLH	1
			Malic_M	1		Dimerisation	1	HLH	1	C2	1	polyprenyl_synt	1
			Vps62	1		NAD_binding_1	1	PRP4	1	Peptidase_C13	1	IF-2	1
			Metallophos	1		DAO	1	HRDC	1	IF-2	1	CytB6-F_Fe-S	1
			zf-CCCH	1		Myb_DNA-bind_6	1	QWRF	1	PGP_phosphatase	1	IlvN	1
			MIF4G	1		COX6B	1	Kinesin	1	IlvN	1	PPR_3	1
			zf-LSD1	1		MORN	1	Response_reg	1	DIOX_N	1	C2	1
			Mito_carr	1		Choline_transpo	1	LEA_1	1	CDC27	1	DIOX_N	1
			AA_permease_2	1		Aha1_N	1	Ribosomal_L21e	1	DnaJ_C	1	Ion_trans_2	1
			MMR_HSR1	1		Cation_efflux	1	Longin	1	Ion_trans	1	RCC1_2	1
						Adenine_glyco	1	RRM_1	1	PP2C	1	Lactamase_B	1
						CAF1	1	LRR_1	1	Ion_trans_2	1	Ribosomal_L2_C	1
						WRKY	1	SBP	1	DUF1298	1	LIAS_N	1
						BTB_2	1	APO_RNA-bind	1	Lactamase_B	1	RRM_2	1
						ABC2_membrane_3	1	Sina	1	Pribosyltran_N	1	zf-RVT	1
						ASL_C	1	MatE	1	LEA_4	1	SBP	1

				AA_permease_C	1	SRP-alpha_N	1	DUF1421	1	ACOX	1
				Alg14	1	Metallophos	1	Lipase_3	1	SNARE	1
				2OG-FeII_Oxy	1	Tmemb_14	1	RCC1_2	1	LRRNT_2	1
				zf-NOSIP	1	Methyltransf_10	1	Chloroa_b-bind	1	SOUL	1
				WD40	1	Transposase_24	1	Ribosomal_L13	1	Amidase	1
				MMR_HSR1	1	Methyltransf_29	1	Mad3_BUB1_I	1	Band_7	1
				Acyltransferase	1	Cation_ATPase	1	Ribosomal_S14	1	BES1_N	1
				MIF4G	1	Mic1	1	14-3-3	1	TPR_1	1
				Abhydrolase_1	1	UCH	1	SBF	1	Methyltransf_11	1
				Mic1	1	zf-UDP	1	Citrate_synt	1	TraB	1
				AAA	1	WD40	1	SCAMP	1	zf-MYND	1
				Methyltransf_29	1	mTERF	1	CorA	1	Trehalose_PPase	1
				AA_permease_2	1	WRKY	1	SNARE	1	Cellulase	1
				Metallophos	1	Myb_DNA-bind_3	1	CS	1	UBD	1
				zf-UDP	1	Yippee-Mis18	1	Solute_trans_a	1	MtN3_slv	1
				MatE	1	Myb_DNA-bind_6	1	Methyltransf_23	1	UQ_con	1
						zf-A20	1	SUI1	1	CDC27	1
						Myb_DNA-binding	1	MtN3_slv	1	WHIM1	1
						zf-C3HC4_2	1	TCP	1	Citrate_synt	1
						NAD_Gly3P_dh_C	1	Cyclin_C	1	DUF2781	1
						zf-CCCH_2	1	ThylakoidFormat	1	NB-LRR	1
						NAF	1	CytB6-F_Fe-S	1	zf-HC5HC2H	1
						zf-RanBP	1	DUF588	1	NHL	1
						BTB_2	1	NHL	1	NifU	1
						NEMP	1	TraB	1	Methyltransf_23	1
						MIF4G	1	NifU	1	Lipase_3	1
								DUF604	1		

							NTP_transferase	1	
							UBD	1	
							OPA3	1	
							DUF688	1	
							Acyl-CoA_dh_M	1	
							WRKY	1	
							PA	1	
							DUF92	1	
							PAM2	1	
							zf-B_box	1	
							PAN_2	1	
							zf-HC5HC2H	1	
							DAGK_cat	1	
							ECH_2	1	
							PCO_ADO	1	
TOTAL	31	93	118	71	140	149		218	189

4.3.2. Identificación de transposones, regiones repetitivas y genes relacionados con elementos transponibles.

En la salida del *repeatmasker* de los elementos repetitivos, fueron separados y limpiados aquellas regiones de baja complejidad como homopolímeros y microsatélites y representaron el 10,16% de elementos transponibles de las regiones de interés, lo cual equivale al 0,3% de secuencias repetitivas de todo el genoma, donde es conocido que en uno de los parentales de *C. arabica*, *C. canephora* es de aproximadamente el 50% de los datos. Todas las regiones presentaron como la familia de transposones más representados la familia LTR variando su frecuencia del 59 al 73%, seguida de LINES, SINES, RC y algunos que no fueron clasificables con la herramienta y la base de datos utilizada (*RepeatMasker* y *RepBase*). Dentro de los LTRs, las familias más frecuentes fueron los Gypsy, seguido de la familia Copia. Como ejemplo está el caso de YLD2, que para el caso del subgenoma *C. canephora*, Gypsy tiene 618 secuencias anotadas y para el caso de la familia Copia tiene 177 (Figuras 14, 15 y 16), lo cual está en concordancia con los reportes de elementos transponibles en café (Dias *et al*., 2015; Guyot *et al.*, 2016).

La búsqueda realizada mediante el uso de palabras clave, sólo arrojó resultados para transposasa en la región PH5 E[a] y con transcriptasa reversa en la región PH5 C[a] y YLD2 E[a].

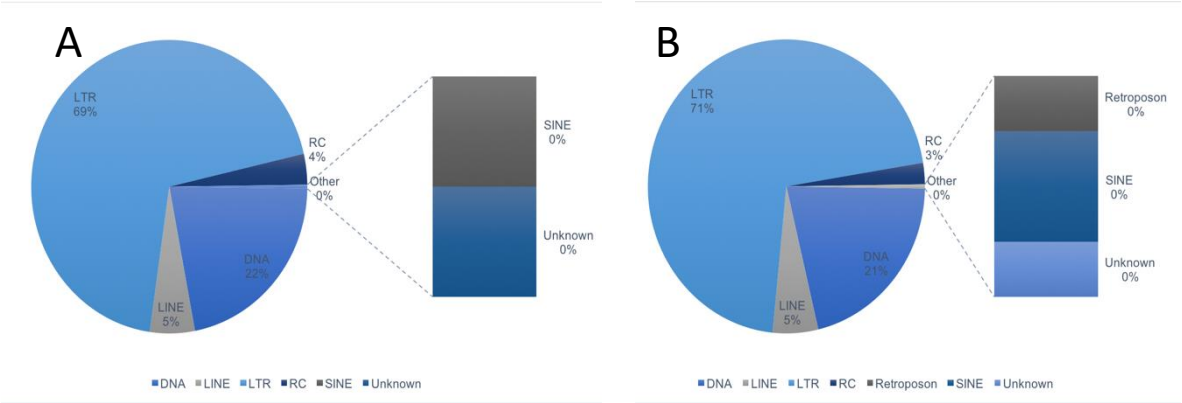


Figura 17. Elementos transponibles para el tamaño del Grano. (A). BS4 C[a]. (B). BS4 E[a].

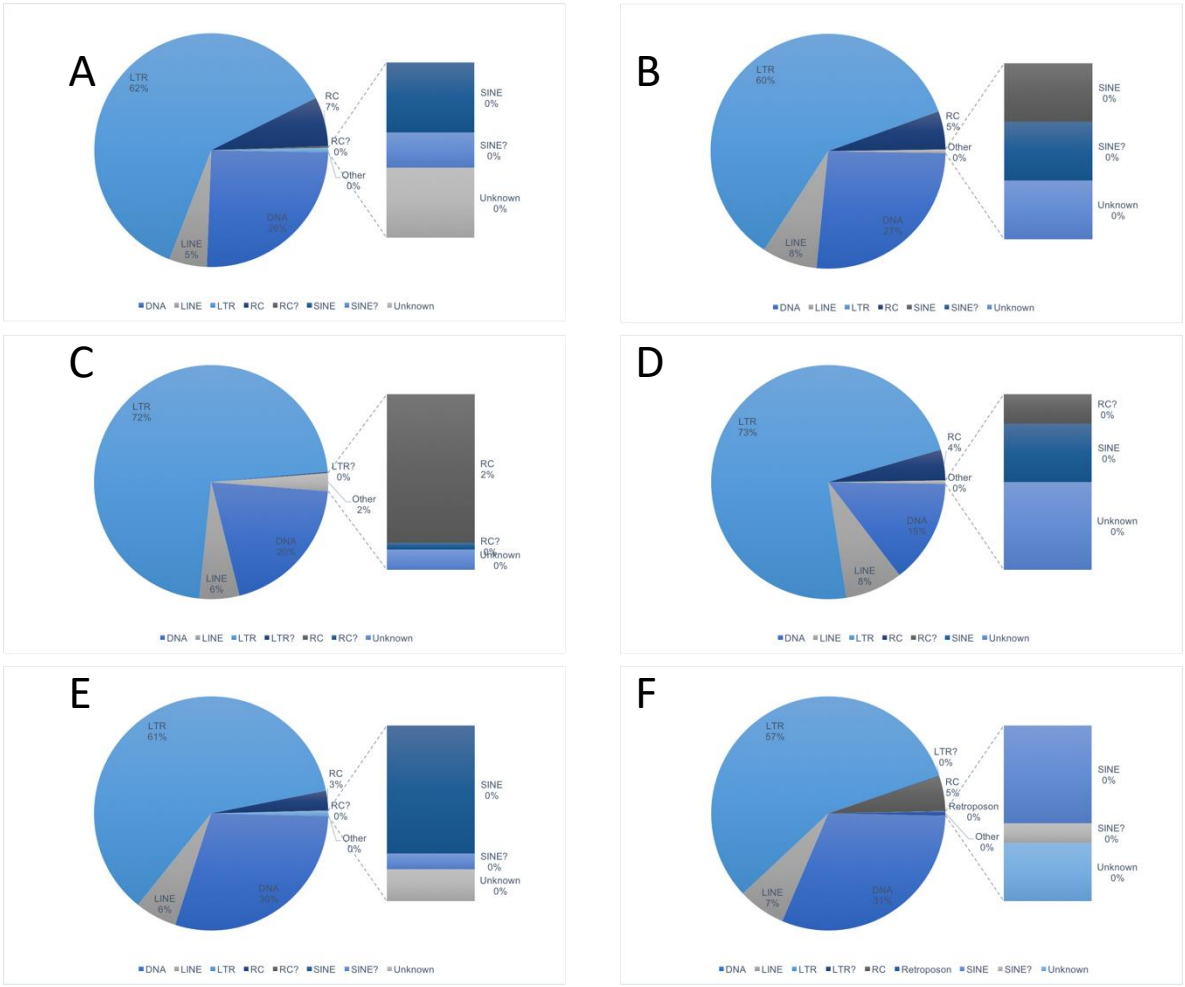


Figura 18. Elementos transponibles para las regiones asociadas a altura de la planta. (A). PH1 C^{a} (B). PH1 E^{a}. (C). PH4 C^{a}. (D). PH4 E^{a}. (E). PH5 C^{a}. (F). PH5 E^{a}.

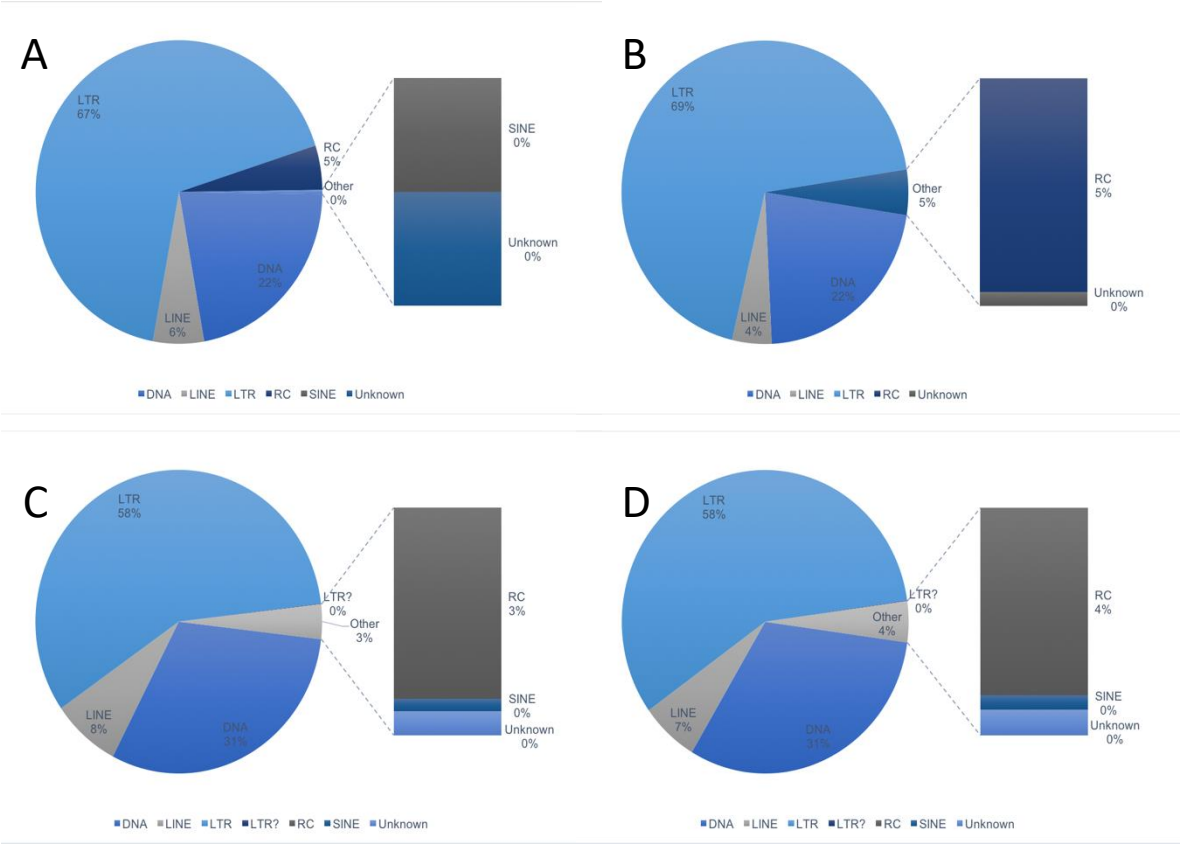


Figura 19. Elementos transponibles para las regiones asociadas a producción. (A). YLD2 C^{a}. (B). YLD2 E^{a}. (C). YLD11 C^{a}. (D). YLD11 E^{a}.

4.3.3. Genómica comparativa

A pesar de no ser uno de los objetivos principales del trabajo se quizo observar si se encontraban mutaciones segmentales en las regiones comparando los subgenomas y el genoma de referencia. BS4 y PH1 muestra regiones muy conservadas con algunas inversiones de regiones muy pequeñas. YLD2, tienen grandes regiones con una alta identidad y apenas pequeños cambios donde se observa una pequeña inversión. Esto contrario a PH4 y PH5, donde se observa una gran cantidad de regiones y se aprecia una gran cantidad de inversiones (Anexo 2). Esto se identifica con claridad cuando se hacen las comparaciones de los dos subgenomas con el genoma de referencia, demostrando una alta cantidad de regiones que aparentemente no están en el mismo orden (Anexo1). Sin embargo, al observar el *dot plot* de la región comparando cada uno de los subgenomas con el genoma de referencia, se denota que existe el mismo orden general de la región al no apreciarse inversiones a esta escala (Figura 15), lo cual fue confirmado al correrse el mismo análisis de los BACs y de los *contigs* ensamblados usando MUMmer donde se confirmó que los *reads* se encuentran en el mismo sentido de la secuencia del genoma de referencia (Datos no mostrados).

Para terminar, al compararse los resultados del presente estudio con el QTL de C. canephora, se puede observar que los grupos de ligamiento tiene algunos de los rasgos y presentan una correlación, lo cual implicaría que estos podrían estar en el mismo grupo de ligación, como por ejemplo en LG B, tamaño del fruto en los años de evaluación 2004 y 2005 con ácidos cafeico-quínicos en 2005. En LG I, colocalización entre amargor, acidez (en 2006), contenido de sacarosa (2005), contenido de 3-CQA, 4- CQA y 5-FQA (2003 y 2005), contenido de cafeína (2005) y rendimiento (2006). En LG J, tamaño del fruto (2006) con ácidos dicaffelquinic (2005) (Leroy, De Bellis, Legnate, Kananura, Gonzales, Pereira, Andrade, Charmetant, Montagnon, Cubry, Marraccini, Pot, & de Kochko, 2011b).

Para *C. arabica*, esto se observó para el grupo de ligamiento 4, que comparte los rasgos BS4 y PH4, los cuales son sobrelapantes, lo cual podría implicar la correlación de los dos rasgos. Estos fueron los únicos que estuvieron presentes en el mismo grupo de ligamiento. En los casos restantes los grupos de ligamiento fueron asociados a un rasgo en particular para cada uno de ellos. YLD2, YLD11, PH1 y PH5.

En *C. canephora* se observaron correlaciones positivas significativas pero bajas entre el rendimiento acumulativo (CY_2002 / 6), el tamaño del fruto (0.21) y el contenido de sacarosa (0.15). El rendimiento acumulado también se correlacionó significativamente con los rasgos organolépticos favorables, como la acidez, el sabor y la nota global (0,19 a 0,29). Los rasgos tecnológicos (tamaño del fruto y tasa) presentaron correlaciones altamente significativas con 3-CQA (0.15 a 0.17) y con 5-FQA (correlaciones negativas de −0.17 a −0.22). Finalmente, las correlaciones entre sabor, acidez y la nota global, con respecto a los rasgos organolépticos, y el contenido de sacarosa y 5-CQA, entre los rasgos bioquímicos, fueron alta y positivamente significativas. Se observó una correlación negativa entre el sabor y la acidez con el contenido de trigonelina y 5-FQA. La amargura se correlacionó negativamente con la sacarosa (−0,25) y se relacionó positivamente con el contenido de cafeína (0,20). En LG B, tamaño de grano en 2004 y 2005 con ácidos cafeico-quínicos en 2005. En LG I, colocalización entre amargor, acidez (en 2006), contenido de sacarosa (2005), contenido de 3-CQA, 4- CQA y 5-FQA (2003 y 2005), contenido de cafeína (2005) y rendimiento (2006). Y en LG J, tamaño del fruto (2006) con ácidos dicaffelquinico (2005) (Leroy, De Bellis, Legnate, Kananura, Gonzales, Pereira, Andrade, Charmetant, Montagnon, Cubry, Marraccini, Pot, & de Kochko, 2011b)**.**

Finalmente, en *C. arabica* los genes asociados a producción y altura de la planta están relacionados con genes involucrados en la actividad celular y acumulación de biomasa, así como enzimas encargadas de la reproducción y de crecimiento. Los dos componentes importantes de los genes encontrados dentro de las regiones de interés están relacionados con la resistencia y la defensa de las plantas como los genes con dominio NB-ARC y LRR, así como la presencia de muchas enzimas P450, que están relacionadas con las especies reactivas de oxígeno e involucradas en procesos de defensa a distancias genéticas <a 30cM y con variantes asociados a dichos genes.

5. Conclusiones generales

5.1. Contribuciones de la Tesis

Se encontraron genes relacionados a resistencia y defensa de la planta como P-quinasas y genes con dominios NB-ARC en todas las regiones asociadas a QTLs.

El 0,34% del genoma corresponde a las regiones asociadas a los rasgos de interés conteniendo genes únicos para cada subgenoma. Para la región BS4 fueron identificados 31 genes para el subgenoma E^a, para PH1 93 para C^a y 118 para E^a para PH5 140 genes en C^a y 149 en E^a y para YLD2 218 en C^a y 189 en E^a.

Las regiones repetitivas corresponden al 0,3% del total del genoma. Con la familia LTR como la más representativa con un 59 al 73% del total de las secuencias repetitivas identificadas en las secuencias asociadas a los rasgos evaluados.

En su mayoría los genes que se encuentran en las regiones asociadas a QTLs son factores de transcripción, y asociados a metabolismo secundario.

La WGS es el camino más simple y económico para comprender la variación estructural de una región sencilla del genoma.

Este trabajo aporta información sobre cuáles son las secuencias que contienen los marcadores asociados a las regiones de interés, y que pueden convertirse en marcadores usados en programas de mejoramiento genético, usándose para el mapeo fino, o para centrarse en ellos para llevar a cabo una genotipificación dirigida para su utilización en el mejoramiento asistido por marcadores, haciendo énfasis en los genes encontrados y que son únicos para cada subgenoma.

5.2. Impactos Potenciales de la Tesis

Apoyo a los investigadores de Cenicafé en el conocimiento de las regiones de interés, y los genes que están probablemente relacionados con estas características.

El ahorro de dinero en las técnicas de secuenciación, pues obtener el genoma completo de calidad por medio de la estrategia WGS es mucho más económico que la secuenciación *targeted sequencing*, y con unos resultados de mejor calidad.

Disminución del tiempo de mejoramiento para la obtención de nuevas variedades, ya que el proceso se demora alrededor de 25 años. Con esta caracterización y con un trabajo posterior en el desarrollo de herramientas de mapeo fino se pueda aportar a la meta de desarrollo de una nueva variedad de café cada 10 a 15 años utilizando mejoramiento asistido por marcadores, o mediante índices genómicos.

Para los agricultores significaría que los nuevos descubrimientos realizados en cuanto a genes y de cómo están estos distribuidos en los subgenomas, llegarían en un tiempo más corto a ellos en el campo, y para el centro de investigación se traduciría en un ahorro, ya que los costos de desarrollo y extensión de una nueva variedad disminuirían drásticamente.

5.3. Recomendaciones y trabajos futuros

Corregir el mapa genético mediante generación de nuevos datos con nuevas herramientas que permita tener una mayor certeza de los grupos de ligamiento, que permita obtener una línea base sólida sobre la cual hacer este tipo de análisis fino que dependen de la ubicación de los marcadores y la certeza del tamaño de la región.

Hacer una genotipificación de los marcadores identificados dentro de la región para establecer su utilidad en un a población para la selección de materiales a utilizar en el mejoramiento genético, y se puedan usar en un futuro como marcadores que puedan apoyar la selección asistida por marcadores.

Es necesario realizar estudios de transcriptómica y epigenómica para establecer subgenoma o alelos activos, así mismo, los genes que están relacionados a las regiones de producción, altura de la planta y tamaño del grano, deben ser probados experimentalmente para establecer su aporte al rasgo cuantitativo.

Como el genoma de *Coffea arabica* fue reportado en el genbank, se recomienda comparar los resultados obtenidos en este trabajo con la secuencia de la base de datos pública ya que esta fue ingresada a refseq, base de datos con alto nivel de curación, y así observar si existen algunos errores en el ensamblaje.

6. Referencias bibliográficas

Akaffou, D. S., Konate, I., Sié, R. S., Crop, V. P. J. O., 2014. (n.d.). Flowering phenology and yield-related traits in an interspecific cross between'Coffea pseudozanguebariae'Bridson and “C. canephora” Pierre. *Search.Informit.com.Au*

Ferrão, L. F. V., Ferrão, R. G., Ferrão, M. A. G., Fonseca, A., Carbonetto, P., Stephens, M., & Garcia, A. A. F. (2018). Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. *Heredity*, 1–15. http://doi.org/10.1038/s41437-018-0105-y

Gartner, G. A. L., McCouch, S. R., & Moncada, M. D. P. (2013). A genetic map of an interspecific diploid pseudo testcross population of coffee. *Euphytica*, *192*(2), 305–323. http://doi.org/10.1007/s10681-013-0926-y

Lepelley, M., Mahesh, V., McCarthy, J., Rigoreau, M., Crouzillat, D., Chabrillange, N., et al. (2012). Characterization, high-resolution mapping and differential expression of three homologous PAL genes in Coffea canephora Pierre (Rubiaceae). *Planta*, *236*(1), 313–326. http://doi.org/10.1007/s00425-012-1613-2

Leroy, T., De Bellis, F., Legnate, H., Kananura, E., Gonzales, G., Pereira, L. F., et al. (2011). Improving the quality of African robustas: QTLs for yield- and quality-related traits in Coffea canephora. *Tree Genetics & Genomes*, *7*(4), 781–798. http://doi.org/10.1007/s11295-011-0374-6

Pestana, K. N., Capucho, A. S., Caixeta, E. T., de Almeida, D. P., Zambolim, E. M., Cruz, C. D., et al. (2015). Inheritance study and linkage mapping of resistance loci to Hemileia vastatrix in Híbrido de Timor UFV 443-03. *Tree Genetics & Genomes*, *11*(4), 890–13. http://doi.org/10.1007/s11295-015-0903-9

Priyono, & PUTRANTO, R. A. (2016). Genetic mapping studies in Coffea sp using molecular marker methods Studi peta genetik pada Coffea sp. menggunakan metode penanda molekuler. *E-Journal Menara Perkebunan*, *83*(2). http://doi.org/10.22302/iribb.jur.mp.v83i2.6

Romero, G., Vásquez, L. M., Lashermes, P., & Herrera, J. C. (2013). Identification of a major QTL for adult plant resistance to coffee leaf rust (Hemileia vastatrix) in the natural Timor hybrid (Coffea arabicax C. canephora). *Plant Breeding*, *133*(1), 121–129. http://doi.org/10.1111/pbr.12127

Akaffou, D. S., Konate, I., Sié, R. S., Crop, V. P. J. O., 2014. (n.d.). Flowering phenology and yield-related traits in an interspecific cross between'Coffea pseudozanguebariae'Bridson and“C. canephora”Pierre. *Search.Informit.com.Au*

.

Anthony, F., Diniz, L. E. C., Combes, M.-C., & Lashermes, P. (2010). Adaptive radiation in Coffea subgenus Coffea L. (Rubiaceae) in Africa and Madagascar. *Plant Systematics and Evolution*, *285*(1-2), 51–64. http://doi.org/10.1007/s00606-009-0255-8

Assmann, S. M. (2013). Natural Variation in Abiotic Stress and Climate Change Responses in Arabidopsis: Implications for Twenty-First-Century Agriculture. *International Journal of Plant Sciences*, *174*(1), 3–26. http://doi.org/10.1086/667798

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 1–6. http://doi.org/10.1186/s13100-015-0041-9

Bolger, M. E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B., & Mayer, K. F. (2014). Plant genome sequencing — applications for crop improvement. *Current Opinion in Biotechnology*, *26*, 31–37. http://doi.org/10.1016/j.copbio.2013.08.019

Bolger, M., Schwacke, R., Gundlach, H., Schmutzer, T., Chen, J., Arend, D., et al. (2017). From Plant Genomes to Phenotypes. *Journal of Biotechnology*, 1–18. http://doi.org/10.1016/j.jbiotec.2017.06.003

Borrill, P., Adamski, N., & Uauy, C. (2015). Genomics as the key to unlocking the polyploid potential of wheat. *New Phytologist*, *208*(4), 1008–1022. http://doi.org/10.1111/nph.13533

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., et al. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiology*, *164*(2), 513–524. http://doi.org/10.1104/pp.113.230144

Cenci, A., Combes, M.-C., & Lashermes, P. (2012). Genome evolution in diploid and tetraploid Coffea species as revealed by comparative analysis of orthologous genome segments. *Plant Molecular Biology*, *78*(1-2), 135–145. http://doi.org/10.1007/s11103-011-9852-3

Chen, M., Presting, G., & Plant, W. B. T. (2002). An integrated physical and genetic map of the rice genome. *Am Soc Plant Biol*
. http://doi.org/10.1105/tpc.010485

Chevin, L.-M., Lande, R., & Mace, G. M. (2010). Adaptation, Plasticity, and Extinction in a Changing Environment: Towards a Predictive Theory. *PLOS Biology*, *8*(4), e1000357–8. http://doi.org/10.1371/journal.pbio.1000357

Crops, A. R. J. O. P., 2008. (n.d.). Speciation of Coffea arabica: implications for genetic improvement.

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research*, *14*(7), 1394–1403. http://doi.org/10.1101/gr.2289704

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, *345*(6201), 1181–1184. http://doi.org/10.1126/science.1255274

Dereeper, A., Bocs, S., Rouard, M., Guignon, V., Ravel, S., Tranchant-Dubreuil, C., et al. (2015). The coffee genome hub: a resource for coffee genomes. *Nucleic Acids Research*, *43*(D1), D1028–D1035. http://doi.org/10.1093/nar/gku1108

Dias, E. S., Hatt, C., Hamon, S., Hamon, P., Rigoreau, M., Crouzillat, D., et al. (2015). Large distribution and high sequence identity of a Copia-type retrotransposon in angiosperm families. *Plant Molecular Biology*, *89*(1), 83–97. http://doi.org/10.1007/s11103-015-0352-8

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., et al. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, *488*(7410), 213–217. http://doi.org/10.1038/nature11241

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042. http://doi.org/10.1111/eva.12178

El-Soda, M., Malosetti, M., Zwaan, B. J., Koornneef, M., & Aarts, M. G. M. (2014). Genotype × environment interaction QTL mapping in plants: lessons from Arabidopsis. *Trends in Plant Science*, 1–10. http://doi.org/10.1016/j.tplants.2014.01.001

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology and Evolution*, *29*(1), 51–63. http://doi.org/10.1016/j.tree.2013.09.008

Fain, S. J., Quiñones, M., Álvarez-Berríos, N. L., Parés-Ramos, I. K., & Gould, W. A. (2017). Climate change and coffee: assessing vulnerability by modeling future climate suitability in the Caribbean island of Puerto Rico, 1–12. http://doi.org/10.1007/s10584-017-1949-5

FAO. (2015). Statistical Pocket Book of the Food And Agricultural Organization for the United Nations, 1–198.

Ferrão, L. F. V., Ferrão, R. G., Ferrão, M. A. G., Fonseca, A., Carbonetto, P., Stephens, M., & Garcia, A. A. F. (2018). Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. *Heredity*, 1–15. http://doi.org/10.1038/s41437-018-0105-y

Gartner, G. A. L., McCouch, S. R., & Moncada, M. D. P. (2013). A genetic map of an interspecific diploid pseudo testcross population of coffee. *Euphytica*, *192*(2), 305–323. http://doi.org/10.1007/s10681-013-0926-y

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, *27*(2), 182–189. http://doi.org/10.1038/nbt.1523

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next- generation sequencing technologies. *Nature Publishing Group*, *17*(6), 333–351. http://doi.org/10.1038/nrg.2016.49

Großkinsky, D. K., Svensgaard, J., Christensen, S., & Roitsch, T. (2015). Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap. *Journal of Experimental Botany*, *66*(18), 5429–5440. http://doi.org/10.1093/jxb/erv345

Guyot, R., Darré, T., Dupeyron, M., Kochko, A., Hamon, S., Couturon, E., et al. (2016). Partial sequencing reveals the transposable element composition of Coffea genomes and provides evidence for distinct evolutionary stories. *Molecular Genetics and Genomics*, *291*(5), 1979–1990. http://doi.org/10.1007/s00438-016-1235-7

Hamon, P., Grover, C. E., Davis, A. P., Rakotomalala, J.-J., Raharimalala, N. E., Albert, V. A., et al. (2017). Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (Coffea) and insights into the evolution of caffeine content in its species: GBS coffee phylogeny and the evolution of caffeine content. *Molecular Phylogenetics and Evolution*, *109*, 351–361. http://doi.org/10.1016/j.ympev.2017.02.009

Harushima, Y., Yano, M., Shomura, A., Sato, M., Shimano, T., Kuboki, Y., et al. (1998). A High-Density Rice Genetic Linkage Map with 2275 Markers Using a Single F2 Population. *Genetics*, *148*(1), 479–494. http://doi.org/10.1073/pnas.90.17.7980

Hoang, N. V., Furtado, A., Mason, P. J., Marquardt, A., Kasirajan, L., Thirugnanasambandam, P. P., et al. (2017). A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing, 1–22. http://doi.org/10.1186/s12864-017-3757-8

Huang, X., & Han, B. (2014). Natural Variations and Genome-Wide Association Studies in Crop Plants. *Annual Review of Plant Biology*, *65*(1), 531–551. http://doi.org/10.1146/annurev-arplant-050213-035715

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. http://doi.org/10.1093/bioinformatics/btu031

Jorgensen, T. H. (2012). The effect of environmental heterogeneity on RPW8-mediated resistance to powdery mildews in Arabidopsis thaliana. *Annals of Botany*, *109*(4), 833–842. http://doi.org/10.1093/aob/mcr320

Juenger, T. E. (2013). Natural variation and genetic constraints on drought tolerance. *Current Opinion in Plant Biology*, *16*(3), 274–281. http://doi.org/10.1016/j.pbi.2013.02.001

Kang, M. S. (1997). Using Genotype-by-Environment Interaction for Crop Cultivar Development. *Advances in Agronomy* (pp. 199–252). http://doi.org/10.1016/s0065-2113(08)60569-6

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, *12*(4), 656–664. http://doi.org/10.1101/gr.229202

Krumsiek, J., Arnold, R., & Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, *23*(8), 1026–1028. http://doi.org/10.1093/bioinformatics/btm039

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*(2), R12–9. http://doi.org/10.1186/gb-2004-5-2-r12

Lacaze, X., Hayes, P. M., & Korol, A. (2008). Genetics of phenotypic plasticity: QTL analysis in barley, Hordeum vulgare. *Heredity*, *102*(2), 163–173. http://doi.org/10.1038/hdy.2008.76

Lashermes, P., Hueber, Y., Combes, M. C., Severac, D., & Dereeper, A. (2016). Inter-genomic DNA exchanges and homeologous gene silencing shaped the nascent allopolyploid coffee genome (Coffea arabica L.). *G Genes, Genomes, Genetics*. http://doi.org/10.1534/g3.116.030858/-/DC1

Lehti-Shiu, M. D., & Shiu, S.-H. (2012). Diversity, classification and function of the plant protein kinase superfamily. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1602), 2619–2639. http://doi.org/10.1098/rstb.2012.0003

Lepelley, M., Mahesh, V., McCarthy, J., Rigoreau, M., Crouzillat, D., Chabrillange, N., et al. (2012). Characterization, high-resolution mapping and differential expression of three homologous PAL genes in Coffea canephora Pierre (Rubiaceae). *Planta*, *236*(1), 313–326. http://doi.org/10.1007/s00425-012-1613-2

Leroy, T., De Bellis, F., Legnate, H., Kananura, E., Gonzales, G., Pereira, L. F., Andrade, A. C., Charmetant, P., Montagnon, C., Cubry, P., Marraccini, P., Pot, D., & de Kochko, A. (2011a). Improving the quality of African robustas: QTLs for yield- and quality-related traits in Coffea canephora. *Tree Genetics & Genomes*, *7*(4), 781–798. http://doi.org/10.1007/s11295-011-0374-6

Leroy, T., De Bellis, F., Legnate, H., Kananura, E., Gonzales, G., Pereira, L. F., Andrade, A. C., Charmetant, P., Montagnon, C., Cubry, P., Marraccini, P., Pot, D., & de Kochko, A. (2011b). Improving the quality of African robustas: QTLs for yield- and quality-

related traits in Coffea canephora. *Tree Genetics & Genomes*, *7*(4), 781–798. http://doi.org/10.1007/s11295-011-0374-6
Li, C., Lin, F., An, D., Wang, W., & Huang, R. (2018). Genome Sequencing and Assembly by Long Reads in Plants. *Genes*, *9*(1), 6–14. http://doi.org/10.3390/genes9010006
Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., et al. (2015). Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. *Nature Biotechnology*, *33*(5), 524–530. http://doi.org/10.1038/nbt.3208
Ling, H.-Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., et al. (2013). Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature*, *496*(7443), 87–90. http://doi.org/10.1038/nature11997
Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265–279. http://doi.org/10.1016/j.gpb.2016.05.004
Mackay, T. F. C. (2004). The genetic architecture of quantitative traits: lessons from Drosophila. *Current Opinion in Genetics & Development*, *14*(3), 253–257. http://doi.org/10.1016/j.gde.2004.04.003
Manna, S. (2015). An overview of pentatricopeptide repeat proteins and their applications. *Biochimie*, *113*(C), 93–99. http://doi.org/10.1016/j.biochi.2015.04.004
Mason, A. S. (2014). Challenges of Genotyping Polyploid Species. In K. Wang (Ed.), *Agrobacterium Protocols* (Vol. 1245, pp. 161–168). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4939-1966-6_12
McClure, K. A., Sawler, J., Gardner, K. M., Money, D., & Myles, S. (2014). Genomics: A potential panacea for the perennial problem. *American Journal of Botany*, *101*(10), 1780–1790. http://doi.org/10.3732/ajb.1400143
McCouch, S., Baute, G. J., Bradeen, J., Bramel, P., Bretting, P. K., Buckler, E., et al. (2013). Agriculture: Feeding the future. *Nature*, *499*(7456), 23–24. http://doi.org/10.1038/499023a
McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., et al. (2009). Genetic Properties of the Maize Nested Association Mapping Population. *Science*, *325*(5941), 737–740. http://doi.org/10.1126/science.1174320
Michael, T. P. (2014). Plant genome size variation: bloating and purging DNA. *Briefings in Functional Genomics*, *13*(4), 308–317. http://doi.org/10.1093/bfgp/elu005
Michael, T. P., & Jackson, S. (2013). The First 50 Plant Genomes. *The Plant Genome Journal*, *6*(2), 0–7. http://doi.org/10.3835/plantgenome2013.03.0001in
Michael, T. P., & VanBuren, R. (2015). Progress, challenges and the future of crop genomes. *Current Opinion in Plant Biology*, *24*, 71–81. http://doi.org/10.1016/j.pbi.2015.02.002
Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., et al. (2018). High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature Communications*, 1–8. http://doi.org/10.1038/s41467-018-03016-2
Ming, R. (2015). Assembling allopolyploid genomes: no longer formidable, 1–4. http://doi.org/10.1186/s13059-015-0585-5
Mishra, A. K., Muthamilarasan, M., Khan, Y., Parida, S. K., & Prasad, M. (2014). Genome-Wide Investigation and Expression Analyses of WD40 Protein Family in the Model Plant Foxtail Millet (Setaria italica L.). *PLoS ONE*, *9*(1), e86852–13. http://doi.org/10.1371/journal.pone.0086852

Mishra, M. K., & Slater, A. (2012). Recent Advances in the Genetic Transformation of Coffee. *Biotechnology Research International*, *2012*(31), 1–17. http://doi.org/10.1155/2012/580857

Moncada, M. D. P., Tovar, E., Montoya, J. C., González, A., Spindel, J., & McCouch, S. (2016). A genetic linkage map of coffee (Coffea arabica L.) and QTL for yield, plant height, and bean size. *Tree Genetics & Genomes*, *12*(1), 5–17. http://doi.org/10.1007/s11295-015-0927-1

Motta, L. B., Soares, T. C. B., Ferrão, M. A. G., Caixeta, E. T., Lorenzoni, R. M., & Souza Neto, J. D. de. (2014). Molecular characterization of arabica and Conilon coffee plants genotypes by SSR and ISSR markers. *Brazilian Archives of Biology and Technology*, *57*(5), 728–735. http://doi.org/10.1590/S1516-8913201402071

Paran, I., & Zamir, D. (2003). Quantitative traits in plants: beyond the QTL. *Trends in Genetics*. http://doi.org/10.1016/S0168-9525(03)00117-3

Pestana, K. N., Capucho, A. S., Caixeta, E. T., de Almeida, D. P., Zambolim, E. M., Cruz, C. D., et al. (2015). Inheritance study and linkage mapping of resistance loci to Hemileia vastatrix in Híbrido de Timor UFV 443-03. *Tree Genetics & Genomes*, *11*(4), 890–13. http://doi.org/10.1007/s11295-015-0903-9

Priyono, & PUTRANTO, R. A. (2016). Genetic mapping studies in Coffea sp using molecular marker methods Studi peta genetik pada Coffea sp. menggunakan metode penanda molekuler. *E-Journal Menara Perkebunan*, *83*(2). http://doi.org/10.22302/iribb.jur.mp.v83i2.6

Renny-Byfield, S., & Wendel, J. F. (2014). Doubling down on genomes: Polyploidy and crop plants. *American Journal of Botany*, *101*(10), 1711–1725. http://doi.org/10.3732/ajb.1400119

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. http://doi.org/10.1016/j.gpb.2015.08.002

Ribas, A. F., Pereira, L. F. P., & Vieira, L. G. E. (2006). Genetic transformation of coffee. *Brazilian Journal of Plant Physiology*, *18*(1), 83–94. http://doi.org/10.1590/S1677-04202006000100007

Rissman, A. I., Mau, B., Biehl, B. S., Darling, A. E., Glasner, J. D., & Perna, N. T. (2009). Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics*, *25*(16), 2071–2073. http://doi.org/10.1093/bioinformatics/btp356

Rivera Calderón, A. L. (2015). Importancia del gen sh3 y su relación con la resistencia a la enfermedad roya del cafeto hemileia vastatrix bajo la modalidad de compilación bibliográfica. *Instname:Universidad Nacional Abierta Y a Distancia*.

Rojas, A. M., Fuentes, G., Rausell, A., & Valencia, A. (2012). The Ras protein superfamily: Evolutionary tree and role of conserved amino acids. *The Journal of Cell Biology*, *196*(2), 189–201. http://doi.org/10.1083/jcb.201103008

Romero, G., Vásquez, L. M., Lashermes, P., & Herrera, J. C. (2013). Identification of a major QTL for adult plant resistance to coffee leaf rust (Hemileia vastatrix) in the natural Timor hybrid (Coffea arabicax C. canephora). *Plant Breeding*, *133*(1), 121–129. http://doi.org/10.1111/pbr.12127

Sattler, M. C., Carvalho, C. R., & Clarindo, W. R. (2015a). The polyploidy and its key role in plant breeding. *Planta*, 1–16. http://doi.org/10.1007/s00425-015-2450-x

Sattler, M. C., Carvalho, C. R., & Clarindo, W. R. (2015b). The polyploidy and its key role in plant breeding. *Planta*, 1–16. http://doi.org/10.1007/s00425-015-2450-x

Schatz, M. C., Maron, L. G., Stein, J. C., Wences, A. H., Gurtowski, J., Biggers, E., et al. (2014). Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biology*, *15*(11), 1631–16. http://doi.org/10.1186/s13059-014-0506-z

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, *326*(5956), 1112–1115. http://doi.org/10.1126/science.1178534

Schuler, G. (1997). Sequence Mapping by Electronic PCR, 1–11.

Shan, Q., Wang, Y., Li, J., & Gao, C. (2014). Genome editing in rice and wheat using the CRISPR/Cas system. *Nature Protocols*, *9*(10), 2395–2410. http://doi.org/10.1038/nprot.2014.157

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., et al. (2010). The genome of woodland strawberry (Fragaria vesca). *Nature Genetics*, *43*(2), 109–116. http://doi.org/10.1038/ng.740

Soderlund, C., Longden, I., & Bioinformatics, R. M. (1997). FPC: a system for building contigs from restriction fingerprinted clones. *Cabios*.

Sommer, D. D., Delcher, A. L., Salzberg, S. L., & Pop, M. (2007). Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics*, *8*(1), 64–11. http://doi.org/10.1186/1471-2105-8-64

Tang, H. (2017). Disentangling a polyploid genome. *Nature Plants*, 1–2. http://doi.org/10.1038/s41477-017-0001-0

Tarailo-Graovac, M., & Chen, N. (2002). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences (Vol. 12, pp. 1269–14). Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.org/10.1002/0471250953.bi0410s25

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., et al. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetics*, *43*(2), 159–162. http://doi.org/10.1038/ng.746

Tilgner, H., Grubert, F., Sharon, D., & Snyder, M. P. (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(27), 9869–9874. http://doi.org/10.1073/pnas.1400447111

Tran, H. T., Lee, L. S., Furtado, A., Smyth, H., & Henry, R. J. (2016). Advances in genomics for the improvement of quality in coffee. *Journal of the Science of Food and Agriculture*, *96*(10), 3300–3312. http://doi.org/10.1002/jsfa.7692

Varshney, R. K., Terauchi, R., & McCouch, S. R. (2014). Harvesting the Promising Fruits of Genomics: Applying Genome Sequencing Technologies to Crop Breeding. *PLOS Biology*, *12*(6), e1001883–8. http://doi.org/10.1371/journal.pbio.1001883

Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., et al. (2013). The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics*, *45*(5), 487–494. http://doi.org/10.1038/ng.2586

Visendi, P., Berkman, P. J., Hayashi, S., Golicz, A. A., Bayer, P. E., Ruperao, P., et al. (2016). An efficient approach to BAC based assembly of complex genomes. *Plant Methods*, 1–9. http://doi.org/10.1186/s13007-016-0107-9

Weigel, D. (2012). Natural Variation in Arabidopsis: From Molecular Genetics to Ecological Genomics. *Plant Physiology*, *158*(1), 2–22. http://doi.org/10.1104/pp.111.189845

Wu, G. A., Prochnik, S., Jenkins, J., Salse, J., Hellsten, U., Murat, F., et al. (2014). Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nature Biotechnology*, 1–8. http://doi.org/10.1038/nbt.2906

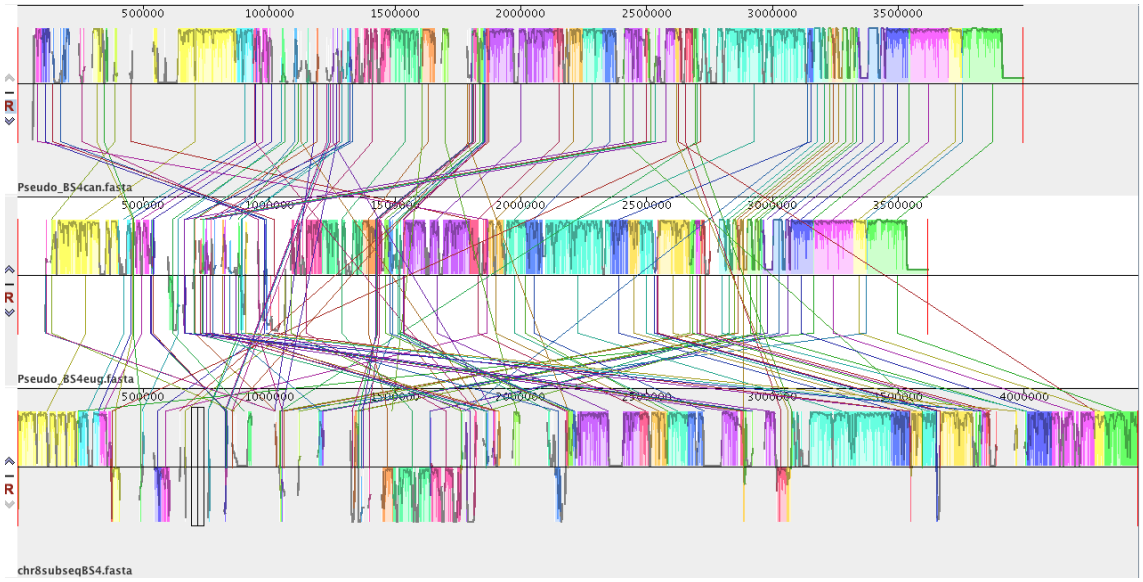
Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S., et al. (2013). The genome of the pear (Pyrus bretschneideri Rehd.). *Genome Research*, *23*(2), 396–408. http://doi.org/10.1101/gr.144311.112

Zhang, J., Chen, L.-L., Xing, F., Kudrna, D. A., Yao, W., Copetti, D., et al. (2016). Extensive sequence divergence between the reference genomes of two elite indicarice varieties Zhenshan 97 and Minghui 63. *Proceedings of the National Academy of Sciences*, *113*(35), E5163–E5171. http://doi.org/10.1073/pnas.1611012113

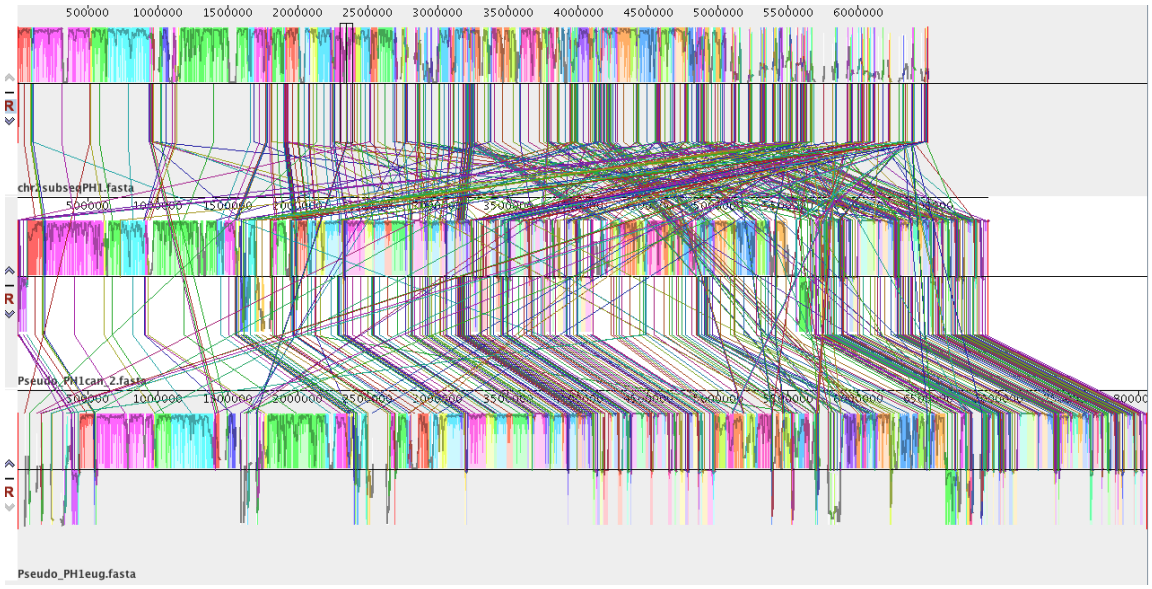
Zimin, A. V., Puiu, D., Hall, R., Kingan, S., Clavijo, B. J., & Salzberg, S. L. (2017). The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *GigaScience*, *6*(11), 1–7. http://doi.org/10.1093/gigascience/gix097

ANEXOS

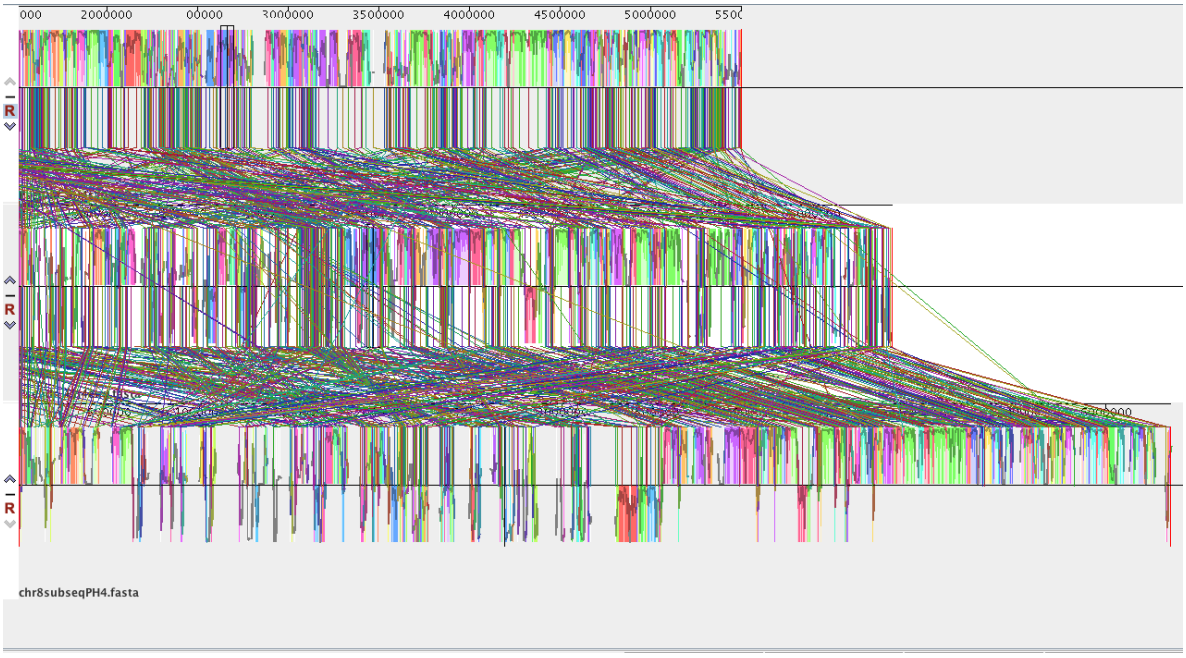
Anexo I. Comparación de las secuencias de los subgenomas de los distintos rasgos con su región correspondiente en *C. canephora*.



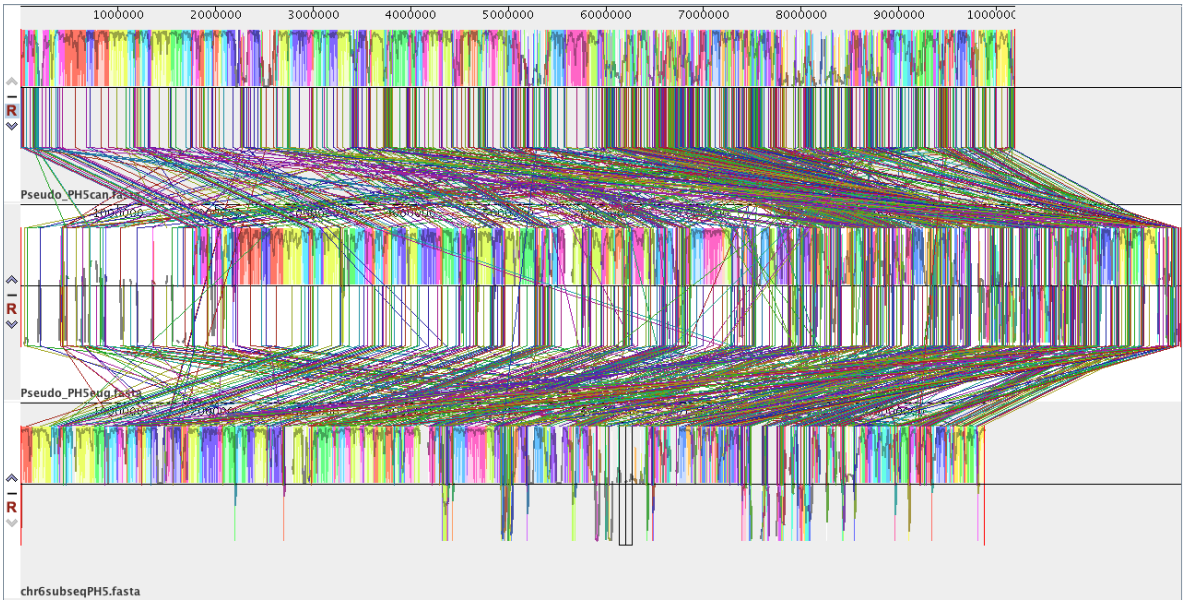
Rasgo tamaño del grano en grupo de ligación 4 - BS4.



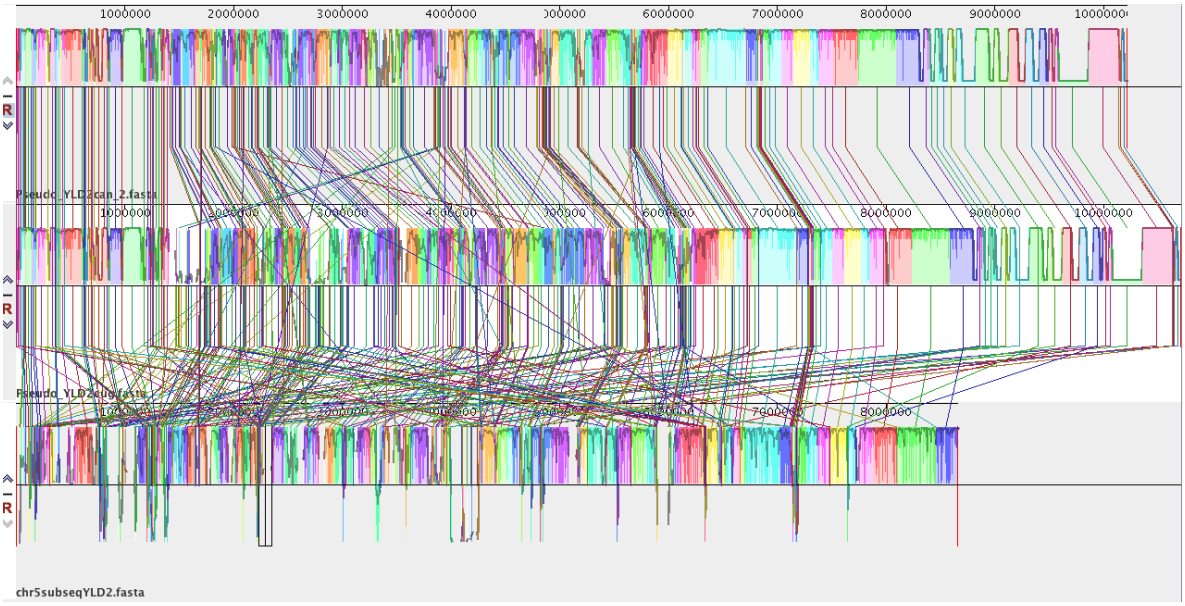
Rasgo Altura de la planta en grupo de ligación 1 - PH1.



Rasgo Altura de la planta en grupo de ligación 4 - PH4.

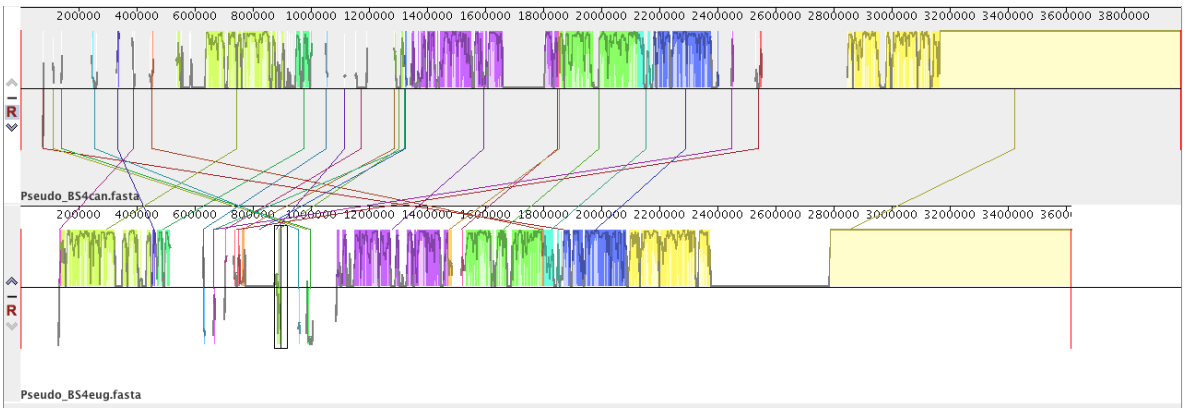


Rasgo Altura de la planta en grupo de ligación 5 - PH5.

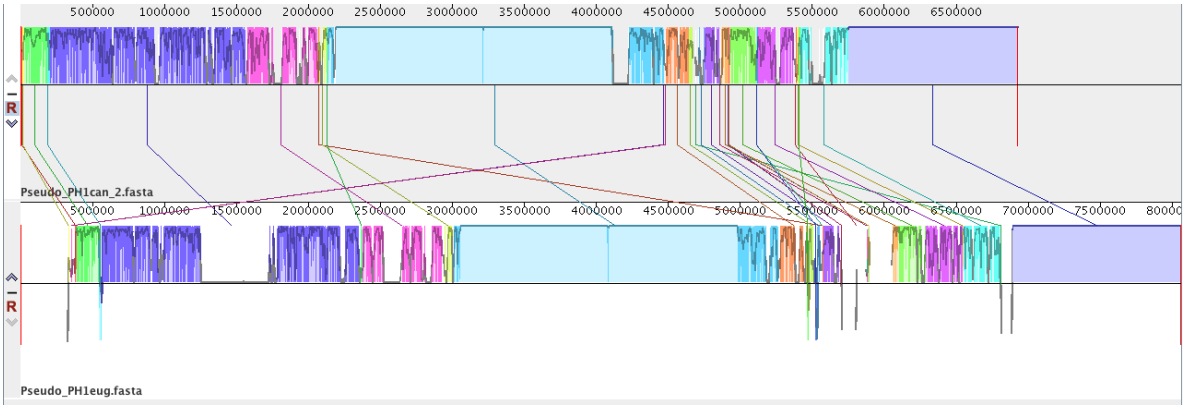


Rasgo Producción en grupo de ligación 2 - YLD2.

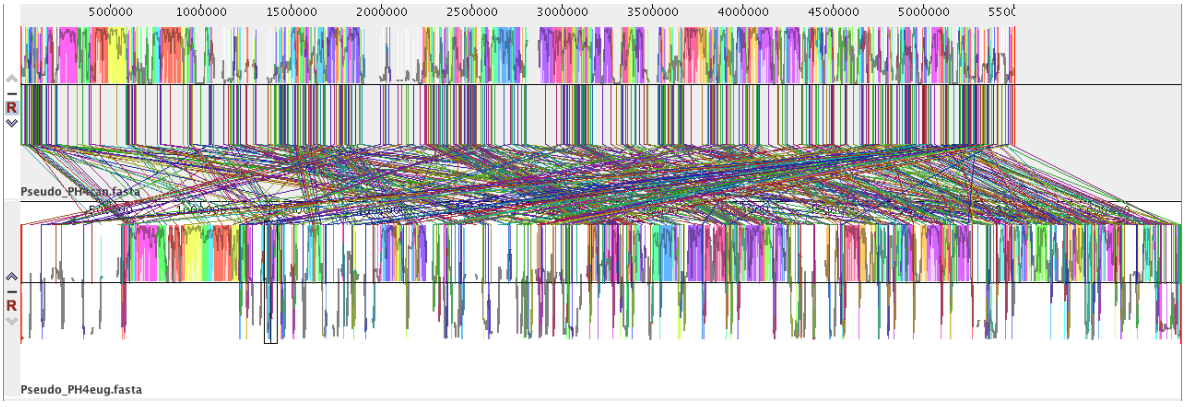
Anexo II. Comparación de las secuencias de los subgenomas de los distintos rasgos.



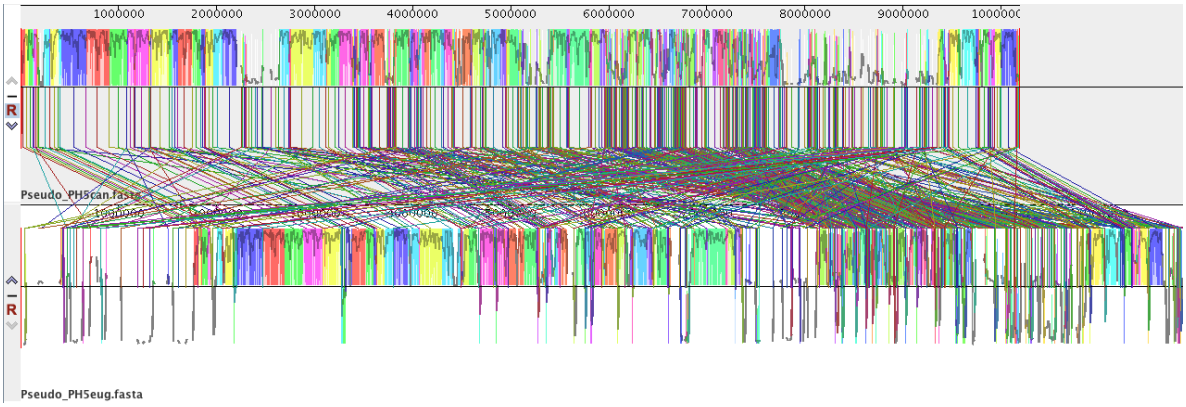
Rasgo tamaño del grano en grupo de ligación 4 - BS4.



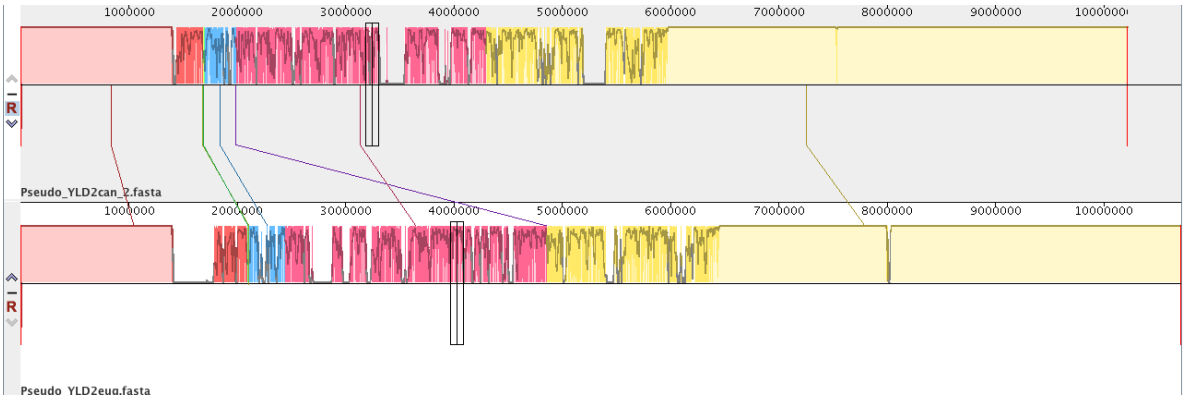
Rasgo Altura de la planta en grupo de ligación 1 – PH1.



Rasgo Altura de la planta en grupo de ligación 4 – PH4.



Rasgo Altura de la planta en grupo de ligación 5 – PH5.



Rasgo Producción en grupo de ligación 2 - YLD2.

Anexo III. Anotación de los genes en las regiones de interés según bases de datos PFAM.

Tabla S 1. Genes Región BS4eug.

GENES	NO. DE GENES	DESCRIPCIÓN
LRR_8	8	Leucine rich repeat
BURP	6	BURP domain
PKINASE	5	Protein kinase domain
DUF247	3	Plant protein of unknown function
PKINASE_TYR	3	Protein tyrosine kinase
GLYCO_HYDRO_32 C	2	Glycosyl hydrolases family 32 C terminal
HYDROLASE_4	2	Serine aminopeptidase, S33
P450	2	Cytochrome P450
WRKY	2	WRKY DNA -binding domain
ZIP	2	ZIP Zinc transporter
ACP_SYN_III_C	1	3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III C terminal
ADH_N	1	Alcohol dehydrogenase GroES-like domain
AMINOTRAN_1_2	1	Aminotransferase class I and II
AMINOTRAN_4	1	Amino-transferase class IV
B2-ADAPT-APP_C	1	Beta2-adaptin appendage, C-terminal sub-domain
BAC_GLOBIN	1	Bacterial-like globin
BRANCH	1	Core-2/I-Branching enzyme
DEAD	1	DEAD/DEAH box helicase
DNA_POL_A	1	DNA polymerase family A
DNA_POL_A_EXO1	1	3'-5' exonuclease
DUF632	1	Protein of unknown function (DUF632)
DUF778	1	Protein of unknown function (DUF778)
DUF810	1	Plant family of unknown function (DUF810)
DUF814	1	Domain of unknown function (DUF814)
ECH_1	1	Enoyl-CoA hydratase/isomerase
EDR1	1	Ethylene-responsive protein kinase Le-CTR1
ENOD93	1	Early nodulin 93 ENOD93 protein
ENTH	1	ENTH domain
EPIMERASE	1	NAD dependent epimerase/dehydratase family
FA_DESATURASE	1	Fatty acid desaturase
FAD_BINDING_4	1	FAD binding domain
FAR1	1	FAR1 DNA-binding domain

FERROCHELATASE	1	Ferrochelatase
FH2	1	Formin Homology 2 Domain
FMO-LIKE	1	Flavin-binding monooxygenase-like
GALACTOSYL_T	1	Galactosyltransferase
GAT	1	GAT domain
GATA	1	GATA zinc finger
GDC-P	1	Glycine cleavage system P-protein
GTP_EFTU	1	Elongation factor Tu GTP binding domain
HIST_DEACETYL	1	Histone deacetylase domain
HYDROLASE_3	1	haloacid dehalogenase-like hydrolase
IQ	1	IQ calmodulin-binding motif
KH_1	1	KH domain
LRR_6	1	Leucine Rich repeat
MALIC_M	1	Malic enzyme, NAD binding domain
MEKHLA	1	MEKHLA domain
MITO_CARR	1	Mitochondrial carrier protein
MMR_HSR1	1	50S ribosome-binding GTPase
MYB_DNA-BIND_6	1	Myb-like DNA-binding domain
NA_H_EXCHANGER	1	Sodium/hydrogen exchanger family
NDR	1	Ndr family
NUCLEOPOR_NUP85	1	Nup85 Nucleoporin
OXIDORED_Q6	1	NADH ubiquinone oxidoreductase, 20 Kd subunit
PAF67	1	RNA polymerase I-associated factor PAF67
PEPTIDASE_C13	1	Peptidase C13 family
PEPTIDASE_M41	1	Peptidase family M41
PEPTIDASE_S10	1	Serine carboxypeptidase
PFK	1	Phosphofructokinase
PITH	1	PITH domain
PNP_UDP_1	1	Phosphorylase superfamily
POX	1	Associated with HOX
PP2C	1	Protein phosphatase 2C
PPR_2	1	PPR repeat family
PROFILIN	1	Profilin
PROT_INHIB_II	1	Potato type II proteinase inhibitor family
RCC1	1	Regulator of chromosome condensation (RCC1) repeat
RHOMBOID	1	Rhomboid family
RIBOSOMAL_L28	1	Ribosomal L28 family
RIBOSOMAL_L5E	1	Ribosomal large subunit proteins 60S L5, and 50S L18
RPE65	1	Retinal pigment epithelial membrane protein
RRM_3	1	RNA binding motif

SAWADEE	1	SAWADEE domain
TAUE	1	Sulfite exporter TauE/SafE
TOCOPHEROL_CYCL	1	Tocopherol cyclase
TRNA_INT_ENDO_N	1	tRNA intron endonuclease, N-terminal domain
TRNA-SYNT_1B	1	tRNA synthetases class I (W and Y)
UQ_CON	1	Ubiquitin-conjugating enzyme
YIF1	1	YIF1
YIPPEE-MIS18	1	Yippee zinc-binding/DNA-binding /Mis18, centromere assembly
ZF-C2H2_JAZ	1	Zinc-finger double-stranded RNA-binding
ZF-RING_2	1	Ring finger domain
TOTAL	107	

Tabla S 2. Genes Región PH1can.

GENES	NO. DE GENES	DESCRIPCIÓN
P450	9	Cytochrome P450
EPIMERASE	6	NAD dependent epimerase/dehydratase family
GLYCO_HYDRO_1	4	Glycosyl hydrolase family 1
PPR_2	4	PPR repeat family
STR_SYNTH	4	Strictosidine synthase
GST_N	3	Glutathione S-transferase, N-terminal domain
DNAJ_C	2	DnaJ C terminal domain
PKINASE	2	Protein kinase domain
PROPEPTIDE_C1	2	Peptidase family C1 propeptide
RAS	2	Ras family
RRM_1	2	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
TRANSFERASE	2	Transferase family
UDPGP	2	UTP--glucose-1-phosphate uridylyltransferase
3BETA_HSD	1	3-beta hydroxysteroid dehydrogenase/isomerase family
AA_PERMEASE_2	1	Amino acid permease
AATF-CHE1	1	Apoptosis antagonizing transcription factor
ALG6_ALG8	1	ALG6, ALG8 glycosyltransferase family
AMINO_OXIDASE	1	Flavin containing amine oxidoreductase
AMP-BINDING_C	1	AMP-binding enzyme C-terminal domain
ARM	1	Armadillo/beta-catenin-like repeat
ATP-SYNT_DE_N	1	ATP synthase, Delta/Epsilon chain, beta-sandwich domain
CITRATE_SYNT	1	Citrate synthase, C-terminal domain
CK_II_BETA	1	Casein kinase II regulatory subunit

CND1_N	1	non-SMC mitotic condensation complex subunit 1, N-term
CU_BIND_LIKE	1	Plastocyanin-like domain
CYCLIN	1	Cyclin
CYS_MET_META_PP	1	Cys/Met metabolism PLP-dependent enzyme
CYT-B5	1	Cytochrome b5-like Heme/Steroid binding domain
DCMP_CYT_DEAM_1	1	Cytidine and deoxycytidylate deaminase zinc-binding region
DNA_TOPOISOIV	1	DNA gyrase/topoisomerase IV, subunit A
DUF1005	1	Protein of unknown function (DUF1005)
DUF1639	1	Protein of unknown function (DUF1639)
DUF4506	1	Domain of unknown function (DUF4506)
DUF630	1	Protein of unknown function (DUF630)
DUF642	1	Protein of unknown function (DUF642)
DYW_DEAMINASE	1	DYW family of nucleic acid deaminases
ECH_2	1	Enoyl-CoA hydratase/isomerase
ELMO_CED12	1	ELMO/CED-12 family
ENT	1	ENT domain
EXOSTOSIN	1	Exostosin family
FKBP_C	1	FKBP-type peptidyl-prolyl cis-trans isomerase
FKS1_DOM1	1	1,3-beta-glucan synthase subunit FKS1, domain-1
G-PATCH	1	G-patch domain
GLYCO_HYDRO_20	1	Glycosyl hydrolase family 20, catalytic domain
GLYCO_TRANSF_22	1	Alg9-like mannosyltransferase family
GLYCOHYDRO_20B2	1	beta-acetyl hexosaminidase like
HAD_2	1	Haloacid dehalogenase-like hydrolase
HALZ	1	Homeobox associated leucine zipper
HD-ZIP_N	1	HD-ZIP protein N terminus
ILVD_EDD	1	Dehydratase family
KH_1	1	KH domain
LCM	1	Leucine carboxyl methyltransferase
LEA_1	1	Late embryogenesis abundant (LEA) group 1
LIN37	1	LIN37
LSM	1	LSM domain
MA3	1	MA3 domain
METALLOPHOS	1	Calcineurin-like phosphoesterase
MITO_CARR	1	Mitochondrial carrier protein
MOB_SYNTH_C	1	Molybdenum Cofactor Synthesis C
MORN	1	MORN repeat
N_ASN_AMIDOHYD	1	Protein N-terminal asparagine amidohydrolase

NAGLU_C	1	Alpha-N-acetylglucosaminidase (NAGLU) C-terminal domain
NB-ARC	1	NB-ARC domain
NDK	1	Nucleoside diphosphate kinase
NEPROSIN_AP	1	Neprosin activation peptide
NIR_SIR_FERR	1	Nitrite/Sulfite reductase ferredoxin-like half domain
NLE	1	NLE (NUC135) domain
O-FUCT	1	GDP-fucose protein O-fucosyltransferase
OBERON_CC	1	Coiled-coil region of Oberon
OHCU_DECARBOX	1	OHCU decarboxylase
OMPDECASE	1	Orotidine 5'-phosphate decarboxylase / HUMPS family
PALM_THIOEST	1	Palmitoyl protein thioesterase
PHD	1	PHD-finger
PKINASE_TYR	1	Protein tyrosine kinase
PLANT_NMP1	1	Plant nuclear matrix protein 1 (NMP1)
PLUS-3	1	Plus-3 domain
PP-BINDING	1	Phosphopantetheine attachment site
PPR_3	1	Pentatricopeptide repeat domain
PTR2	1	POT family
RIBOSOMAL_L1	1	Ribosomal protein L1p/L10e family
RIBOSOMAL_L3	1	Ribosomal protein L3
RIBOSOMAL_L37AE	1	Ribosomal L37ae protein family
RIBOSOMAL_S7E	1	Ribosomal protein S7e
RPE65	1	Retinal pigment epithelial membrane protein
RWD	1	RWD domain
SCAI	1	Protein SCAI
SEC23_BS	1	Sec23/Sec24 beta-sandwich domain
SEC63	1	Sec63 Brl domain
SIGMA70_R2	1	Sigma-70 region 2
STEROL_MT_C	1	Sterol methyltransferase C-terminal
STM1_N	1	Stm1
TIM21	1	TIM21
TP_METHYLASE	1	Tetrapyrrole (Corrin/Porphyrin) Methylases
TPR_2	1	Tetratricopeptide repeat
TRNA_SYNT_2F	1	Glycyl-tRNA synthetase beta subunit
TRYP_ALPHA_AMYL	1	Protease inhibitor/seed storage/LTP family
UBIA	1	UbiA prenyltransferase family
UBIE_METHYLTRAN	1	ubiE/COQ5 methyltransferase family
UFM1	1	Ubiquitin fold modifier 1 protein
UTP21	1	Utp21 specific WD40 associated putative domain

VPS62	1	Vacuolar protein sorting-associated protein 62
ZF-CCCH_2	1	RNA-binding, Nab2-type zinc finger
ZF-CCCH_3	1	Zinc-finger containing family
ZF-LSD1	1	LSD1 zinc finger
ZF-MET	1	Zinc-finger of C2H2 type
TOTAL	136	

Tabla S 3. Genes Región PH1eug.

GENES	NO. DE GENES	DESCRIPCIÓN
P450	8	Cytochrome P450
EPIMERASE	5	NAD dependent epimerase/dehydratase family
GLYCO_HYDRO_1	5	Glycosyl hydrolase family 1
2OG-FEII_OXY	4	2OG-Fe(II) oxygenase superfamily
HAD_2	3	Haloacid dehalogenase-like hydrolase
PPR_2	3	PPR repeat family
RAS	3	Ras family
STR_SYNTH	3	Strictosidine synthase
DNAJ_C	2	DnaJ C terminal domain
DUF1005	2	Protein of unknown function (DUF1005)
PKINASE	2	Protein kinase domain
PPR_3	2	Pentatricopeptide repeat domain
RRM_1	2	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
TRANSFERASE	2	Transferase family
WAX2_C	2	WAX2 C-terminal domain
ZF-UDP	2	Zinc-binding RING-finger
3BETA_HSD	1	3-beta hydroxysteroid dehydrogenase/isomerase family
AA_PERMEASE_2	1	Amino acid permease
AATF-CHE1	1	Apoptosis antagonizing transcription factor
ABC_MEMBRANE	1	ABC transporter transmembrane region
ADH_ZINC_N	1	Zinc-binding dehydrogenase
ALG6_ALG8	1	ALG6, ALG8 glycosyltransferase family
AMINO_OXIDASE	1	Flavin containing amine oxidoreductase
ATP-SYNT_DE_N	1	ATP synthase, Delta/Epsilon chain, beta-sandwich domain
CBFB_NFYA	1	CCAAT-binding transcription factor (CBF-B/NF-YA) subunit B
CK_II_BETA	1	Casein kinase II regulatory subunit
CND1_N	1	non-SMC mitotic condensation complex subunit 1, N-term
CPN10	1	Chaperonin 10 Kd subunit
CU_BIND_LIKE	1	Plastocyanin-like domain

CYCLIN	1	Cyclin
CYT-B5	1	Cytochrome b5-like Heme/Steroid binding domain
DCMP_CYT_DEAM_1	1	Cytidine and deoxycytidylate deaminase zinc-binding region
DDE_TNP_4	1	DDE superfamily endonuclease
DNA_TOPOISOIV	1	DNA gyrase/topoisomerase IV, subunit A
DUF4477	1	Domain of unknown function (DUF4477)
DUF498	1	Protein of unknown function (DUF498/DUF598)
DUF632	1	Protein of unknown function (DUF632)
DYW_DEAMINASE	1	DYW family of nucleic acid deaminases
ECH_2	1	Enoyl-CoA hydratase/isomerase
EIF2A	1	Eukaryotic translation initiation factor eIF2A
ELMO_CED12	1	ELMO/CED-12 family
ENT	1	ENT domain
EXOSTOSIN	1	Exostosin family
F-BOX-LIKE	1	F-box-like
FER4_7	1	4Fe-4S dicluster domain
FKS1_DOM1	1	1,3-beta-glucan synthase subunit FKS1, domain-1
FMN_RED	1	NADPH-dependent FMN reductase
FRA10AC1	1	Folate-sensitive fragile site protein Fra10Ac1
GED	1	Dynamin GTPase effector domain
GLYCO_HYDRO_20	1	Glycosyl hydrolase family 20, catalytic domain
GLYCO_TRANSF_22	1	Alg9-like mannosyltransferase family
GLYCOHYDRO_20B2	1	beta-acetyl hexosaminidase like
GST_N	1	Glutathione S-transferase, N-terminal domain
HALZ	1	Homeobox associated leucine zipper
HD-ZIP_N	1	HD-ZIP protein N terminus
HELICASE_C	1	Helicase conserved C-terminal domain
HIS_PHOS_1	1	Histidine phosphatase superfamily (branch 1)
HLH	1	Helix-loop-helix DNA-binding domain
ILVD_EDD	1	Dehydratase family
KH_1	1	KH domain
LCM	1	Leucine carboxyl methyltransferase
LEA_1	1	Late embryogenesis abundant (LEA) group 1
LIN37	1	LIN37
LSM	1	LSM domain
MALIC_M	1	Malic enzyme, NAD binding domain
METALLOPHOS	1	Calcineurin-like phosphoesterase
MIF4G	1	MIF4G domain
MITO_CARR	1	Mitochondrial carrier protein

MMR_HSR1	1	50S ribosome-binding GTPase
MOB_SYNTH_C	1	Molybdenum Cofactor Synthesis C
MORN	1	MORN repeat
N_ASN_AMIDOHYD	1	Protein N-terminal asparagine amidohydrolase
NAGLU_C	1	Alpha-N-acetylglucosaminidase (NAGLU) C-terminal domain
NB-ARC	1	NB-ARC domain
NDK	1	Nucleoside diphosphate kinase
NEPROSIN_AP	1	Neprosin activation peptide
NIR_SIR_FERR	1	Nitrite/Sulfite reductase ferredoxin-like half domain
NLE	1	NLE (NUC135) domain
NOG1	1	Nucleolar GTP-binding protein 1 (NOG1)
NSP1_C	1	Nsp1-like C-terminal region
O-FUCT	1	GDP-fucose protein O-fucosyltransferase
OMPDECASE	1	Orotidine 5'-phosphate decarboxylase / HUMPS family
PALM_THIOEST	1	Palmitoyl protein thioesterase
PHD_OBERON	1	PHD - plant homeodomain finger protein
PKINASE_TYR	1	Protein tyrosine kinase
PLANT_NMP1	1	Plant nuclear matrix protein 1 (NMP1)
PP-BINDING	1	Phosphopantetheine attachment site
PRIBOSYL_SYNTH	1	Phosphoribosyl synthetase-associated domain
PROPEPTIDE_C1	1	Peptidase family C1 propeptide
PTR2	1	POT family
RCBX	1	RbcX protein
RIBOSOMAL_L1	1	Ribosomal protein L1p/L10e family
RIBOSOMAL_L18P	1	Ribosomal L18 of archaea, bacteria, mitoch. and chloroplast
RIBOSOMAL_L3	1	Ribosomal protein L3
RIBOSOMAL_L37AE	1	Ribosomal L37ae protein family
RIBOSOMAL_S7E	1	Ribosomal protein S7e
RPE65	1	Retinal pigment epithelial membrane protein
RWD	1	RWD domain
SCAI	1	Protein SCAI
SEC23_BS	1	Sec23/Sec24 beta-sandwich domain
SEC61_BETA	1	Sec61beta family
SEC63	1	Sec63 Brl domain
SIGMA70_R4	1	Sigma-70, region 4
STEROL_MT_C	1	Sterol methyltransferase C-terminal
STM1_N	1	Stm1
TELOMERASE_RBD	1	Telomerase ribonucleoprotein complex - RNA binding domain
TIM21	1	TIM21

TPR_2	1	Tetratricopeptide repeat
TRNA_SYNT_2F	1	Glycyl-tRNA synthetase beta subunit
TRNA-SYNT_2D	1	tRNA synthetases class II core domain (F)
TRYP_ALPHA_AMYL	1	Protease inhibitor/seed storage/LTP family
UBIA	1	UbiA prenyltransferase family
UBIE_METHYLTRAN	1	ubiE/COQ5 methyltransferase family
UFM1	1	Ubiquitin fold modifier 1 protein
UPF0565	1	Uncharacterised protein family UPF0565
VPS62	1	Vacuolar protein sorting-associated protein 62
ZF-CCCH	1	Zinc finger C-x8-C-x5-C-x3-H type (and similar)
ZF-CCCH_2	1	RNA-binding, Nab2-type zinc finger
ZF-LSD1	1	LSD1 zinc finger
TOTAL	153	

Tabla S 4. Genes Región PH4eug.

GENES	NO. DE GENES	DESCRIPCIÓN
DUF1350	4	Protein of unknown function (DUF1350)
ABC_MEMBRANE	2	ABC transporter transmembrane region
ABC_TRAN	2	ABC transporter
LRR_8	2	Leucine rich repeat
NB-ARC	2	NB-ARC domain
RRM_1	2	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
WD40	2	WD domain, G-beta repeat
2OG-FEII_OXY_2	1	2OG-Fe(II) oxygenase superfamily
ABHYDROLASE_3	1	alpha/beta hydrolase fold
ACBP	1	Acyl CoA binding protein
ADAPTIN_N	1	Adaptin N terminal region
AOX	1	Alternative oxidase
ARFGAP	1	Putative GTPase activating protein for Arf
B3	1	B3 DNA binding domain
BAH	1	BAH domain
C1_2	1	C1 domain
CELLULOSE_SYNT	1	Cellulose synthase
COBW_C	1	Cobalamin synthesis protein cobW C-terminal domain
COG6	1	Conserved oligomeric complex COG6
DAD	1	DAD family
DIM1	1	Mitosis protein DIM1

DIOX_N	1	non-haem dioxygenase in morphine synthesis N-terminal
DUF1336	1	Protein of unknown function (DUF1336)
DUF1664	1	Protein of unknown function (DUF1664)
DUF5110	1	Domain of unknown function (DUF5110)
DUF647	1	Vitamin B6 photo-protection and homoeostasis
EAMA	1	EamA-like transporter family
EXO_ENDO_PHOS	1	Endonuclease/Exonuclease/phosphatase family
GLYCO_HYDRO_19	1	Chitinase class I
GLYCO_HYDRO_31	1	Glycosyl hydrolases family 31
GPAT_N	1	Glycerol-3-phosphate acyltransferase N-terminal
GWT1	1	GWT1
HYDROLASE_4	1	Serine aminopeptidase, S33
ISO_DH	1	Isocitrate/isopropylmalate dehydrogenase
JMJN	1	jmjN domain
LSM14	1	Scd6-like Sm domain
MED26	1	TFIIS helical bundle-like domain
MFMR_ASSOC	1	Disordered region downstream of MFMR
MYB_CC_LHEQLE	1	MYB-CC type transfactor, LHEQLE motif
MYB_DNA-BIND_3	1	Myb/SANT-like DNA-binding domain
MYB_DNA-BINDING	1	Myb-like DNA-binding domain
NAD_BINDING_10	1	NAD(P)H-binding
NUCLEOTID_TRANS	1	Nucleotide-diphospho-sugar transferase
PAE	1	Pectinacetylesterase
PAP_FIBRILLIN	1	PAP_fibrillin
PEPTIDASE_M49	1	Peptidase family M49
PHM7_CYT	1	Cytosolic domain of 10TM putative phosphate transporter
PIG-X	1	PIG-X / PBN1
PKINASE_TYR	1	Protein tyrosine kinase
PLATZ	1	PLATZ transcription factor
PPTA	1	Protein prenyltransferase alpha subunit repeat
PTEN_C2	1	C2 domain of PTEN tumour-suppressor protein
RIBOSOMAL_L13E	1	Ribosomal protein L13e
RIBOSOMAL_L35P	1	Ribosomal protein L35
RNA_POL_RPB2_5	1	RNA polymerase Rpb2, domain 5
RNA_POL_RPC34	1	RNA polymerase Rpc34 subunit
RPE65	1	Retinal pigment epithelial membrane protein
RSN1_TM	1	Late exocytosis, associated with Golgi transport
RST	1	RCD1-SRO-TAF4 (RST) plant domain
RWD	1	RWD domain
START	1	START domain

SYNTAXIN_2	1	Syntaxin-like protein
THIF	1	ThiF family
TRANSGLUT_CORE 2	1	Transglutaminase-like superfamily
TREHALOSE_PPASE	1	Trehalose-phosphatase
UCH	1	Ubiquitin carboxyl-terminal hydrolase
XET_C	1	Xyloglucan endo-transglycosylase (XET) C-terminus
ZF-CCHC_6	1	Zinc knuckle
TOTAL	77	

Tabla S 5. Genes Región PH5can.

GENES	NO. DE GENES	DESCRIPCIÓN
PKINASE	13	Protein kinase domain
LRR_8	6	Leucine rich repeat
HELICASE_C	5	Helicase conserved C-terminal domain
DIOX_N	4	non-haem dioxygenase in morphine synthesis N-terminal
NMRA	4	NmrA-like family
PKINASE_TYR	4	Protein tyrosine kinase
PPR_2	4	PPR repeat family
WD40	3	WD domain, G-beta repeat
ACTIN	2	Actin
ADH_SHORT_C2	2	Enoyl-(Acyl carrier protein) reductase
ANF_RECEPTOR	2	Receptor family ligand binding region
ANK_2	2	Ankyrin repeats (3 copies)
CAATP_NAI	2	Ca2+-ATPase N terminal autoinhibitory domain
CAF1C_H4-BD	2	Histone-binding protein RBBP4 or subunit C of CAF1 complex
COX6B	2	Cytochrome oxidase c subunit VIb
DUF1668	2	Protein of unknown function (DUF1668)
EIF3_N	2	eIF3 subunit 6 N terminal domain
MS_CHANNEL	2	Mechanosensitive ion channel
NUC173	2	NUC173 domain
PC-ESTERASE	2	GDSL/SGNH-like Acyl-Esterase family found in Pmr5 and Cas1p
PECTINESTERASE	2	Pectinesterase
PEPTIDASE_M16_C	2	Peptidase M16 inactive domain
PEPTIDASE_M16_M	2	Middle or third domain of peptidase_M16
PEROXIDASE	2	Peroxidase
PP2C	2	Protein phosphatase 2C

RDRP	2	RNA dependent RNA polymerase
RECA	2	recA bacterial DNA recombination protein
RNASE_PH_C	2	3' exoribonuclease family, domain 2
SNF7	2	Snf7
TF_ZN_RIBBON	2	TFIIB zinc-binding
TPT	2	Triose-phosphate Transporter family
TRANSFERASE	2	Transferase family
2OG-FEII_OXY	1	2OG-Fe(II) oxygenase superfamily
5_NUCLEOTID	1	5' nucleotidase family
AA_PERMEASE_2	1	Amino acid permease
AA_PERMEASE_C	1	C-terminus of AA_permease
AAA	1	ATPase family associated with various cellular activities (AAA)
AAA_26	1	AAA domain
ABC2_MEMBRANE_3	1	ABC-2 family transporter protein
ABHYDROLASE_1	1	alpha/beta hydrolase fold
ABI	1	CAAX protease self-immunity
ACETYLTRANSF_1	1	Acetyltransferase (GNAT) family
ACETYLTRANSF_10	1	Acetyltransferase (GNAT) domain
ACP_SYN_III	1	3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III
ACYLTRANSFERASE	1	Acyltransferase
ADENINE_GLYCO	1	Methyladenine glycosylase
ADIP	1	Afadin- and alpha -actinin-Binding
AFG1_ATPASE	1	AFG1-like ATPase
AHA1_N	1	Activator of Hsp90 ATPase, N-terminal
ALDEDH	1	Aldehyde dehydrogenase family
ALDOSE_EPIM	1	Aldose 1-epimerase
ALG14	1	Oligosaccharide biosynthesis protein Alg14 like
ALO	1	D-arabinono-1,4-lactone oxidase
AMP_N	1	Aminopeptidase P, N-terminal domain
ANTH	1	ANTH domain
APO_RNA-BIND	1	APO RNA-binding
ARF	1	ADP-ribosylation factor family
ASL_C	1	Adenylosuccinate lyase C-terminal
ATP-SYNT_G	1	Mitochondrial ATP synthase g subunit
AUXIN_INDUCIBLE	1	Auxin responsive protein
AUXIN_RESP	1	Auxin response factor
BAAT_C	1	BAAT / Acyl-CoA thioester hydrolase C terminal
BRF1	1	Brf1-like TBP-binding domain

BTB_2	1	BTB/POZ domain
CAF1	1	CAF1 family ribonuclease
CALMODULIN_BIND	1	Calmodulin binding protein-like
CARB_ANHYDRASE	1	Eukaryotic-type carbonic anhydrase
CATION_EFFLUX	1	Cation efflux family
CCT	1	CCT motif
CCT_2	1	Divergent CCT motif
CH	1	Calponin homology (CH) domain
CHOLINE_TRANSPO	1	Plasma-membrane choline transporter
CLAT_ADAPTOR_S	1	Clathrin adaptor complex small chain
CLP_N	1	Clp amino terminal domain, pathogenicity island component
CN_HYDROLASE	1	Carbon-nitrogen hydrolase
COQ4	1	Coenzyme Q (ubiquinone) biosynthesis protein Coq4
COQ9	1	COQ9
CPN10	1	Chaperonin 10 Kd subunit
CRAL_TRIO_2	1	Divergent CRAL/TRIO domain
CUPIN_4	1	Cupin superfamily protein
CWFJ_C_1	1	Protein similar to CwfJ C-terminus 1
CYTOKIN-BIND	1	Cytokinin dehydrogenase 1, FAD and cytokinin binding
DAGK_CAT	1	Diacylglycerol kinase catalytic domain
DALR_1	1	DALR anticodon binding domain
DAO	1	FAD dependent oxidoreductase
DBINO	1	DNA-binding domain
DER1	1	Der1-like family
DIL	1	DIL domain
DIMERISATION	1	Dimerisation domain
DNA_POL_E_B	1	DNA polymerase alpha/epsilon subunit B
DNA_POL3_DELTA2	1	DNA polymerase III, delta subunit
DNAJ	1	DnaJ domain
DPBB_1	1	Lytic transglycolase
DS	1	Deoxyhypusine synthase
DSPC	1	Dual specificity phosphatase, catalytic domain
DUF1313	1	Protein of unknown function (DUF1313)
DUF2838	1	Protein of unknown function (DUF2838)
DUF3336	1	Domain of unknown function (DUF3336)
DUF3474	1	Domain of unknown function (DUF3474)
DUF3479	1	Domain of unknown function (DUF3479)
DUF3490	1	Domain of unknown function (DUF3490)

DUF4005	1	Protein of unknown function (DUF4005)
DUF4408	1	Domain of unknown function (DUF4408)
DUF563	1	Protein of unknown function (DUF563)
DUF676	1	Putative serine esterase (DUF676)
DUF788	1	Protein of unknown function (DUF788)
DUF810	1	Plant family of unknown function (DUF810)
DXP_SYNTHASE_N	1	1-deoxy-D-xylulose-5-phosphate synthase
DYW_DEAMINASE	1	DYW family of nucleic acid deaminases
EAMA	1	EamA-like transporter family
EF-HAND_6	1	EF-hand domain
EIF-5A	1	Eukaryotic elongation factor 5A hypusine, DNA-binding OB fold
EMP70	1	Endomembrane protein 70
ENT	1	ENT domain
ESTERASE	1	Putative esterase
EXO_ENDO_PHOS	1	Endonuclease/Exonuclease/phosphatase family
F-BOX	1	F-box domain
FA_DESATURASE	1	Fatty acid desaturase
FA_FANCE	1	Fanconi Anaemia group E protein FANCE
FAD_BINDING_1	1	FAD binding domain
FAD_BINDING_3	1	FAD binding domain
FH2	1	Formin Homology 2 Domain
G_GLU_TRANSPEPT	1	Gamma-glutamyltranspeptidase
GATA	1	GATA zinc finger
GELSOLIN	1	Gelsolin repeat
GLUTAREDOXIN	1	Glutaredoxin
GLYCO_HYDRO_16	1	Glycosyl hydrolases family 16
GLYCO_HYDRO_17	1	Glycosyl hydrolases family 17
GLYCO_HYDRO_28	1	Glycosyl hydrolases family 28
GLYCO_HYDRO_35	1	Glycosyl hydrolases family 35
GLYCO_TRANSF_22	1	Alg9-like mannosyltransferase family
GOLGIN_A5	1	Golgin subfamily A member 5
GPI1	1	N-acetylglucosaminyl transferase component (Gpi1)
HAD_2	1	Haloacid dehalogenase-like hydrolase
HAT1_N	1	Histone acetyl transferase HAT1 N-terminus
HD_3	1	HD domain
HD-ZIP_N	1	HD-ZIP protein N terminus

HEXAPEP	1	Bacterial transferase hexapeptide (six repeats)
HEXOKINASE_2	1	Hexokinase
HLH	1	Helix-loop-helix DNA-binding domain
HLYIII	1	Haemolysin-III related
HSF_DNA-BIND	1	HSF-type DNA-binding
IF-2B	1	Initiation factor 2 subunit family
INHIBITOR_I29	1	Cathepsin propeptide inhibitor domain (I29)
IST1	1	Regulator of Vps4 activity in the MVB pathway
KETOACYL-SYNT_C	1	Beta-ketoacyl synthase, C-terminal domain
KH_2	1	KH domain
LCAT	1	Lecithin:cholesterol acyltransferase
LDH_1_N	1	lactate/malate dehydrogenase, NAD binding domain
LEA_1	1	Late embryogenesis abundant (LEA) group 1
LIPASE_GDSL	1	GDSL-like Lipase/Acylhydrolase
LONGIN	1	Regulated-SNARE-like domain
LRR_6	1	Leucine Rich repeat
LYCOPENE_CYCL	1	Lycopene cyclase protein
MALIC_M	1	Malic enzyme, NAD binding domain
MATE	1	MatE
MBOAT	1	MBOAT, membrane-bound O-acyltransferase family
MED20	1	TATA-binding related factor (TRF) of subunit 20 of Mediator complex
MEM_TRANS	1	Membrane transport protein
METALLOPHOS	1	Calcineurin-like phosphoesterase
METHYLTRANSF_11	1	Methyltransferase domain
METHYLTRANSF_29	1	Putative S-adenosyl-L-methionine-dependent methyltransferase
METHYLTRANSF_3	1	O-methyltransferase
MIC1	1	Colon cancer-associated protein Mic1-like
MIF4G	1	MIF4G domain
MITO_CARR	1	Mitochondrial carrier protein
MMR_HSR1	1	50S ribosome-binding GTPase
MORN	1	MORN repeat
MYB_DNA-BIND_6	1	Myb-like DNA-binding domain
NAD_BINDING_1	1	Oxidoreductase NAD-binding domain
NAD_BINDING_10	1	NAD(P)H-binding
NAD_BINDING_11	1	NAD-binding of NADP-dependent 3-hydroxyisobutyrate dehydrogenase
NAD_BINDING_6	1	Ferric reductase NAD binding domain
NAD_GLY3P_DH_C	1	NAD-dependent glycerol-3-phosphate dehydrogenase C-terminus

NAF	1	NAF domain
NAM	1	No apical meristem (NAM) protein
NAP	1	Nucleosome assembly protein (NAP)
NB-ARC	1	NB-ARC domain
NDR	1	Ndr family
NEMP	1	NEMP family
NIC96	1	Nup93/Nic96
NIF	1	NLI interacting factor-like phosphatase
NMT	1	Myristoyl-CoA:protein N-myristoyltransferase, N-terminal domain
NPR1_LIKE_C	1	NPR1/NIM1 like defence protein C terminal
NST1	1	Salt tolerance down-regulator
NT_GLN_AMIDASE	1	N-terminal glutamine amidase
NT-C2	1	N-terminal C2 in EEIG1 and EHBP1 proteins
NUDIX_2	1	Nucleotide hydrolase
NUDIX_N_2	1	Nudix N-terminal
O-FUCT	1	GDP-fucose protein O-fucosyltransferase
OB_NTP_BIND	1	Oligonucleotide/oligosaccharide-binding (OB)-fold
OXIDORED_FMN	1	NADH:flavin oxidoreductase / NADH oxidase family
P5CR_DIMER	1	Pyrroline-5-carboxylate reductase dimerisation
PAS_9	1	PAS domain
PCI	1	PCI domain
PCRF	1	PCRF domain
PEPTIDASE_M17_N	1	Cytosol aminopeptidase family, N-terminal domain
PEPTIDASE_M20	1	Peptidase family M20/M25/M40
PEPTIDASE_S10	1	Serine carboxypeptidase
PEPTIDASE_S28	1	Serine carboxypeptidase S28
PEPTIDASE_S9	1	Prolyl oligopeptidase family
PFK	1	Phosphofructokinase
PGG	1	Domain of unknown function
PHD_OBERON	1	PHD - plant homeodomain finger protein
PIP5K	1	Phosphatidylinositol-4-phosphate 5-Kinase
PKINASE_C	1	Protein kinase C terminal domain
PMEI	1	Plant invertase/pectin methylesterase inhibitor
PMR5N	1	PMR5 N terminal Domain
PORIN_3	1	Eukaryotic porin
PPR	1	PPR repeat
PRMA	1	Ribosomal protein L11 methyltransferase (PrmA)
PROFILIN	1	Profilin

PROTEASOME_A_N	1	Proteasome subunit A N-terminal signature
PRP4	1	pre-mRNA processing factor 4 (PRP4) like
PUNUT	1	Purine nucleobase transmembrane transport
PUT_PHOSPHATASE	1	Putative Phosphatase
PYR_REDOX_2	1	Pyridine nucleotide-disulphide oxidoreductase
QWRF	1	QWRF family
RED_N	1	RED-like protein N-terminal region
REPA_OB_2	1	Replication protein A OB domain
RIBOSOMAL_60S	1	60s Acidic ribosomal protein
RIBOSOMAL_L21E	1	Ribosomal protein L21e
RNA_POL	1	DNA-dependent RNA polymerase
RRM_1	1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
RVT_3	1	Reverse transcriptase-like
SACCHRP_DH_C	1	Saccharopine dehydrogenase C-terminal domain
SATASE_N	1	Serine acetyltransferase, N-terminal
SEC34	1	Sec34-like family
SECY	1	SecY translocase
SENESCENCE	1	Senescence-associated protein
SINA	1	Seven in absentia protein family
SPC97_SPC98	1	Spc97 / Spc98 family
SRP-ALPHA_N	1	Signal recognition particle, alpha subunit, N-terminal
STR_SYNTH	1	Strictosidine synthase
STT3	1	Oligosaccharyl transferase STT3 subunit
SUCROSE_SYNTH	1	Sucrose synthase
SUI1	1	Translation initiation factor SUI1
SUZ	1	SUZ domain
SYJA_N	1	SacI homology domain
SYNTAXIN-6_N	1	Syntaxin 6, N-terminal
TETRASPANNIN	1	Tetraspanin family
THIOREDOXIN	1	Thioredoxin
TMEMB_14	1	Transmembrane proteins 14C
TPR_12	1	Tetratricopeptide repeat
TPR_19	1	Tetratricopeptide repeat
TPR_8	1	Tetratricopeptide repeat
TREHALOSE_PPASE	1	Trehalose-phosphatase
TRNA_ANTI-CODON	1	OB-fold nucleic acid binding domain
TRYP_ALPHA_AMYL	1	Protease inhibitor/seed storage/LTP family
TTL	1	Tubulin-tyrosine ligase family

U-BOX	1	U-box domain
UBIA	1	UbiA prenyltransferase family
UBIQUITIN	1	Ubiquitin family
UCH	1	Ubiquitin carboxyl-terminal hydrolase
UDENN	1	uDENN domain
UPF0066	1	Uncharacterised protein family UPF0066
VPS53_N	1	Vps53-like, N-terminal
WAX2_C	1	WAX2 C-terminal domain
WRC	1	WRC
WRKY	1	WRKY DNA -binding domain
WW	1	WW domain
XET_C	1	Xyloglucan endo-transglycosylase (XET) C-terminus
YBAB_DNA_BD	1	YbaB/EbfC DNA-binding family
YIP1	1	Yip1 domain
YIPPEE-MIS18	1	Yippee zinc-binding/DNA-binding /Mis18, centromere assembly
YSC84	1	Las17-binding protein actin regulator
ZEIN-BINDING	1	Zein-binding
ZF-C6H2	1	zf-MYND-like zinc finger, mRNA-binding
ZF-CCCH	1	Zinc finger C-x8-C-x5-C-x3-H type (and similar)
ZF-FLZ	1	zinc-finger of the FCS-type, C2-C2
ZF-LSD1	1	LSD1 zinc finger
ZF-MET	1	Zinc-finger of C2H2 type
ZF-NOSIP	1	Zinc-finger of nitric oxide synthase-interacting protein
ZF-UDP	1	Zinc-binding RING-finger
TOTAL	333	

Tabla S 6. Genes Región PH5eug.

GENES	NO. DE GENES	DESCRIPCIÓN
PKINASE	10	Protein kinase domain
PKINASE_TYR	8	Protein tyrosine kinase
LRR_8	7	Leucine rich repeat
DIOX_N	6	non-haem dioxygenase in morphine synthesis N-terminal
ALDEDH	4	Aldehyde dehydrogenase family
ANF_RECEPTOR	4	Receptor family ligand binding region

GLYCO_HYDRO_18	4	Glycosyl hydrolases family 18
HELICASE_C	4	Helicase conserved C-terminal domain
RAS	4	Ras family
ACTIN	3	Actin
ANK_2	3	Ankyrin repeats (3 copies)
EXOSTOSIN	3	Exostosin family
GLYCO_HYDRO_17	3	Glycosyl hydrolases family 17
NB-ARC	3	NB-ARC domain
PPR_2	3	PPR repeat family
RRM_1	3	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
WD40	3	WD domain, G-beta repeat
ZF-RANBP	3	Zn-finger in Ran binding protein and others
AA_PERMEASE_2	2	Amino acid permease
ABHYDROLASE_1	2	alpha/beta hydrolase fold
ACETYLTRANSF_10	2	Acetyltransferase (GNAT) domain
ADH_SHORT_C2	2	Enoyl-(Acyl carrier protein) reductase
CAF1C_H4-BD	2	Histone-binding protein RBBP4 or subunit C of CAF1 complex
CU_BIND_LIKE	2	Plastocyanin-like domain
DYW_DEAMINASE	2	DYW family of nucleic acid deaminases
EIF3_N	2	eIF3 subunit 6 N terminal domain
FAD_BINDING_6	2	Oxidoreductase FAD-binding domain
HEXOKINASE_2	2	Hexokinase
KINESIN	2	Kinesin motor domain
LRRNT_2	2	Leucine rich repeat N-terminal domain
MITO_CARR	2	Mitochondrial carrier protein
MS_CHANNEL	2	Mechanosensitive ion channel
MTERF	2	mTERF
NAF	2	NAF domain
O-FUCT	2	GDP-fucose protein O-fucosyltransferase
P450	2	Cytochrome P450
PECTINESTERASE	2	Pectinesterase
PEPTIDASE_M16_C	2	Peptidase M16 inactive domain
PEPTIDASE_S10	2	Serine carboxypeptidase
PEPTIDASE_S24	2	Peptidase S24-like
PEPTIDASE_S8	2	Subtilase family
PEROXIDASE	2	Peroxidase
PGG	2	Domain of unknown function
PP2C	2	Protein phosphatase 2C

PPR_LONG	2	Pentacotripeptide-repeat region of PRORP
RDRP	2	RNA dependent RNA polymerase
SUGAR_TR	2	Sugar (and other) transporter
SYNTAXIN-5_N	2	Syntaxin-5 N-terminal, Sly1p-binding domain
TETRASPANNIN	2	Tetraspanin family
TRYP_ALPHA_AMYL	2	Protease inhibitor/seed storage/LTP family
TTL	2	Tubulin-tyrosine ligase family
UDPGT	2	UDP-glucoronosyl and UDP-glucosyl transferase
XET_C	2	Xyloglucan endo-transglycosylase (XET) C-terminus
JAS	1	TPL-binding domain in jasmonate signalling
2OG-FEII_OXY	1	2OG-Fe(II) oxygenase superfamily
3HCDH_N	1	3-hydroxyacyl-CoA dehydrogenase, NAD binding domain
5_NUCLEOTID	1	5' nucleotidase family
AAA	1	ATPase family associated with various cellular activities (AAA)
AAA_26	1	AAA domain
ABC_TRAN	1	ABC transporter
ABC_TRANS_N	1	ABC-transporter extracellular N-terminal
ABHYDROLASE_2	1	Phospholipase/Carboxylesterase
ABHYDROLASE_6	1	Alpha/beta hydrolase family
ACETYLTRANSF_1	1	Acetyltransferase (GNAT) family
ACP_SYN_III	1	3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III
ACP_SYN_III_C	1	3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III C terminal
ACYL-ACP_TE	1	Acyl-ACP thioesterase
ADENINE_GLYCO	1	Methyladenine glycosylase
ADH_SHORT	1	short chain dehydrogenase
ADIP	1	Afadin- and alpha -actinin-Binding
AFG1_ATPASE	1	AFG1-like ATPase
AHPC-TSA	1	AhpC/TSA family
AHSA1	1	Activator of Hsp90 ATPase homolog 1-like protein
ALBA	1	Alba
ALDOSE_EPIM	1	Aldose 1-epimerase
ALG14	1	Oligosaccharide biosynthesis protein Alg14 like
ALO	1	D-arabinono-1,4-lactone oxidase
AMINO_OXIDASE	1	Flavin containing amine oxidoreductase
AMINOTRAN_1_2	1	Aminotransferase class I and II
AMP_N	1	Aminopeptidase P, N-terminal domain
AMP-BINDING_C	1	AMP-binding enzyme C-terminal domain
ANAPC4_WD40	1	Anaphase-promoting complex subunit 4 WD40 domain
AP2	1	AP2 domain

APO_RNA-BIND	1	APO RNA-binding
ARF	1	ADP-ribosylation factor family
ATG11	1	Autophagy-related protein 11
ATP_BIND_3	1	PP-loop family
ATP-SYNT_G	1	Mitochondrial ATP synthase g subunit
AUX_IAA	1	AUX/IAA family
AUXIN_INDUCIBLE	1	Auxin responsive protein
AUXIN_RESP	1	Auxin response factor
B56	1	Protein phosphatase 2A regulatory B subunit (B56 family)
BIN3	1	Bicoid-interacting protein 3 (Bin3)
BRANCH	1	Core-2/I-Branching enzyme
BRAP2	1	BRCA1-associated protein 2
BTB_2	1	BTB/POZ domain
CAF1	1	CAF1 family ribonuclease
CALMODULIN_BIND	1	Calmodulin binding protein-like
CAM_BINDING	1	Plant calmodulin-binding domain
CARB_ANHYDRASE	1	Eukaryotic-type carbonic anhydrase
CATION_ATPASE	1	Cation transport ATPase (P-type)
CATION_EFFLUX	1	Cation efflux family
CCT	1	CCT motif
CCT_2	1	Divergent CCT motif
CHOLINE_KINASE	1	Choline/ethanolamine kinase
CHOLINE_TRANSPO	1	Plasma-membrane choline transporter
CN_HYDROLASE	1	Carbon-nitrogen hydrolase
COQ9	1	COQ9
COX6B	1	Cytochrome oxidase c subunit VIb
CPN10	1	Chaperonin 10 Kd subunit
CRAL_TRIO_2	1	Divergent CRAL/TRIO domain
CRCB	1	CrcB-like protein, Camphor Resistance (CrcB)
CRS1_YHBY	1	CRS1 / YhbY (CRM) domain
CSN8_PSD8_EIF3K	1	CSN8/PSMD8/EIF3K family
CUE	1	CUE domain
CWFJ_C_1	1	Protein similar to CwfJ C-terminus 1
CYTOKIN-BIND	1	Cytokinin dehydrogenase 1, FAD and cytokinin binding
DAGK_ACC	1	Diacylglycerol kinase accessory domain
DAGK_CAT	1	Diacylglycerol kinase catalytic domain
DALR_1	1	DALR anticodon binding domain
DAO	1	FAD dependent oxidoreductase

DEAD	1	DEAD/DEAH box helicase
DENN	1	DENN (AEX-3) domain
DIMERISATION	1	Dimerisation domain
DNA_POL3_DELTA2	1	DNA polymerase III, delta subunit
DNAJ	1	DnaJ domain
DPBB_1	1	Lytic transglycolase
DS	1	Deoxyhypusine synthase
DUF1313	1	Protein of unknown function (DUF1313)
DUF1517	1	Protein of unknown function (DUF1517)
DUF1644	1	Protein of unknown function (DUF1644)
DUF1977	1	Domain of unknown function (DUF1977)
DUF2048	1	Abhydrolase domain containing 18
DUF2358	1	Uncharacterized conserved protein (DUF2358)
DUF3119	1	Protein of unknown function (DUF3119)
DUF3336	1	Domain of unknown function (DUF3336)
DUF3474	1	Domain of unknown function (DUF3474)
DUF3490	1	Domain of unknown function (DUF3490)
DUF3548	1	Domain of unknown function (DUF3548)
DUF4005	1	Protein of unknown function (DUF4005)
DUF4187	1	Domain of unknown function (DUF4187)
DUF4339	1	Domain of unknown function (DUF4339)
DUF4413	1	Domain of unknown function (DUF4413)
DUF4743	1	Domain of unknown function (DUF4743)
DUF538	1	Protein of unknown function, DUF538
DUF676	1	Putative serine esterase (DUF676)
DUF773	1	Protein of unknown function (DUF773)
DUF788	1	Protein of unknown function (DUF788)
DUF810	1	Plant family of unknown function (DUF810)
DUF92	1	Integral membrane protein DUF92
DUF936	1	Plant protein of unknown function (DUF936)
ECH_2	1	Enoyl-CoA hydratase/isomerase
EF_TS	1	Elongation factor TS
EF-HAND_5	1	EF hand
EF-HAND_7	1	EF-hand domain pair
EIF-5A	1	Eukaryotic elongation factor 5A hypusine, DNA-binding OB fold
EMP70	1	Endomembrane protein 70
ENT	1	ENT domain
ESTERASE	1	Putative esterase
FA_DESATURASE	1	Fatty acid desaturase

FAD_BINDING_1	1	FAD binding domain
FAD_BINDING_3	1	FAD binding domain
FCF1	1	Fcf1
FGGY_C	1	FGGY family of carbohydrate kinases, C-terminal domain
FIP1	1	Fip1 motif
FPL	1	Uncharacterised conserved protein
FYVE	1	FYVE zinc finger
G_GLU_TRANSPEPT	1	Gamma-glutamyltranspeptidase
GATA	1	GATA zinc finger
GDPD	1	Glycerophosphoryl diester phosphodiesterase family
GLYCO_HYDRO_28	1	Glycosyl hydrolases family 28
GLYCO_HYDRO_35	1	Glycosyl hydrolases family 35
GLYCO_TRANSF_22	1	Alg9-like mannosyltransferase family
GLYCOS_TRANSF_4	1	Glycosyl transferase family 4
GLYOXAL_OXID_N	1	Glyoxal oxidase N-terminus
GLYOXALASE	1	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily
GPI1	1	N-acetylglucosaminyl transferase component (Gpi1)
HAD_2	1	Haloacid dehalogenase-like hydrolase
HAT1_N	1	Histone acetyl transferase HAT1 N-terminus
HD_3	1	HD domain
HD-ZIP_N	1	HD-ZIP protein N terminus
HEAT	1	HEAT repeat
HEME_OXYGENASE	1	Heme oxygenase
HEXAPEP	1	Bacterial transferase hexapeptide (six repeats)
HISTONE_H2A_C	1	C-terminus of histone H2A
HLH	1	Helix-loop-helix DNA-binding domain
HLYIII	1	Haemolysin-III related
HRDC	1	HRDC domain
HYDROLASE	1	haloacid dehalogenase-like hydrolase
HYDROLASE_4	1	Serine aminopeptidase, S33
IF-2B	1	Initiation factor 2 subunit family
IST1	1	Regulator of Vps4 activity in the MVB pathway
KELCH_1	1	Kelch motif
KETOACYL-SYNT_C	1	Beta-ketoacyl synthase, C-terminal domain
KH_2	1	KH domain
LCAT	1	Lecithin:cholesterol acyltransferase

LEA_1	1	Late embryogenesis abundant (LEA) group 1
LIPASE_GDSL	1	GDSL-like Lipase/Acylhydrolase
LISH	1	LisH
LONGIN	1	Regulated-SNARE-like domain
LRR_1	1	Leucine Rich Repeat
LRR_6	1	Leucine Rich repeat
LYCOPENE_CYCL	1	Lycopene cyclase protein
LYSM	1	LysM domain
MALIC_M	1	Malic enzyme, NAD binding domain
MATE	1	MatE
MBD	1	Methyl-CpG binding domain
MED17	1	Subunit 17 of Mediator complex
METALLOPHOS	1	Calcineurin-like phosphoesterase
METHYLTRANSF_10	1	Protein of unknown function (DUF890)
METHYLTRANSF_11	1	Methyltransferase domain
METHYLTRANSF_29	1	Putative S-adenosyl-L-methionine-dependent methyltransferase
METHYLTRANSF_3	1	O-methyltransferase
MIC1	1	Colon cancer-associated protein Mic1-like
MIF4G	1	MIF4G domain
MIOX	1	Myo-inositol oxygenase
MTS	1	Methyltransferase small domain
MUTS_III	1	MutS domain III
MYB_DNA-BIND_3	1	Myb/SANT-like DNA-binding domain
MYB_DNA-BIND_6	1	Myb-like DNA-binding domain
MYB_DNA-BINDING	1	Myb-like DNA-binding domain
MYOSIN_N	1	Myosin N-terminal SH3-like domain
NAD_BINDING_10	1	NAD(P)H-binding
NAD_BINDING_11	1	NAD-binding of NADP-dependent 3-hydroxyisobutyrate dehydrogenase
NAD_BINDING_6	1	Ferric reductase NAD binding domain
NAD_GLY3P_DH_C	1	NAD-dependent glycerol-3-phosphate dehydrogenase C-terminus
NAP	1	Nucleosome assembly protein (NAP)
NDR	1	Ndr family
NEMP	1	NEMP family
NMT	1	Myristoyl-CoA:protein N-myristoyltransferase, N-terminal domain
NPR1_LIKE_C	1	NPR1/NIM1 like defence protein C terminal
NSE4_C	1	Nse4 C-terminal
NUC173	1	NUC173 domain

NUCLEOSIDE_TRAN	1	Nucleoside transporter
NUDIX_2	1	Nucleotide hydrolase
NUDIX_N_2	1	Nudix N-terminal
NUP192	1	Nuclear pore complex scaffold, nucleoporins 186/192/205
OB_NTP_BIND	1	Oligonucleotide/oligosaccharide-binding (OB)-fold
OTU	1	OTU-like cysteine protease
OXIDORED_FMN	1	NADH:flavin oxidoreductase / NADH oxidase family
P5CR_DIMER	1	Pyrroline-5-carboxylate reductase dimerisation
PARA	1	NUBPL iron-transfer P-loop NTPase
PC-ESTERASE	1	GDSL/SGNH-like Acyl-Esterase family found in Pmr5 and Cas1p
PCO_ADO	1	PCO_ADO
PCRF	1	PCRF domain
PDCD7	1	Programmed cell death protein 7
PEPTIDASE_C97	1	PPPDE putative peptidase domain
PEPTIDASE_M17_N	1	Cytosol aminopeptidase family, N-terminal domain
PEPTIDASE_M20	1	Peptidase family M20/M25/M40
PEPTIDASE_S28	1	Serine carboxypeptidase S28
PEPTIDASE_S41	1	Peptidase family S41
PEPTIDASE_S9	1	Prolyl oligopeptidase family
PFK	1	Phosphofructokinase
PHD_OBERON	1	PHD - plant homeodomain finger protein
PIP5K	1	Phosphatidylinositol-4-phosphate 5-Kinase
PKINASE_C	1	Protein kinase C terminal domain
PLDC	1	Phospholipase D Active site motif
PMEI	1	Plant invertase/pectin methylesterase inhibitor
PORIN_3	1	Eukaryotic porin
PPR_1	1	PPR repeat
PROFILIN	1	Profilin
PRONE	1	PRONE (Plant-specific Rop nucleotide exchanger)
PROTEASOME	1	Proteasome subunit
PRP4	1	pre-mRNA processing factor 4 (PRP4) like
PRT_C	1	Plant phosphoribosyltransferase C-terminal
PTR2	1	POT family
PUT_PHOSPHATASE	1	Putative Phosphatase
QLQ	1	QLQ
QWRF	1	QWRF family
RAB5IP	1	Rab5-interacting protein (Rab5ip)
RABGAP-TBC	1	Rab-GTPase-TBC domain

RAFFINOSE_SYN	1	Raffinose synthase or seed imbibition protein Sip1
RECA	1	recA bacterial DNA recombination protein
RECQ_ZN_BIND	1	RecQ zinc-binding
RED_N	1	RED-like protein N-terminal region
REDOXIN	1	Redoxin
REPA_OB_2	1	Replication protein A OB domain
RESPONSE_REG	1	Response regulator receiver domain
RETICULON	1	Reticulon
RIBOSOMAL_60S	1	60s Acidic ribosomal protein
RIBOSOMAL_L21E	1	Ribosomal protein L21e
RIBOSOMAL_L35P	1	Ribosomal protein L35
RIBOSOMAL_L7AE	1	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family
RIBOSOMAL_S30	1	Ribosomal protein S30
RNASE_PH_C	1	3' exoribonuclease family, domain 2
RPOL_N	1	DNA-directed RNA polymerase N-terminal
RRS1	1	Ribosome biogenesis regulatory protein (RRS1)
SAP	1	SAP domain
SATASE_N	1	Serine acetyltransferase, N-terminal
SBF	1	Sodium Bile acid symporter family
SBP	1	SBP domain
SEC34	1	Sec34-like family
SECY	1	SecY translocase
SENESCENCE	1	Senescence-associated protein
SHMT	1	Serine hydroxymethyltransferase
SINA	1	Seven in absentia protein family
SKIP_SNW	1	SKIP/SNW domain
SNAPC_SNAP43	1	Small nuclear RNA activating complex (SNAPc), subunit SNAP43
SNF7	1	Snf7
SPC97_SPC98	1	Spc97 / Spc98 family
SPX	1	SPX domain
SRP-ALPHA_N	1	Signal recognition particle, alpha subunit, N-terminal
STAS	1	STAS domain
STT3	1	Oligosaccharyl transferase STT3 subunit
SUCROSE_SYNTH	1	Sucrose synthase
SUI1	1	Translation initiation factor SUI1
SYNTAXIN-6_N	1	Syntaxin 6, N-terminal
TIM17	1	Tim17/Tim22/Tim23/Pmp24 family
TMEMB_14	1	Transmembrane proteins 14C
TPR_12	1	Tetratricopeptide repeat

TPR_19	1	Tetratricopeptide repeat
TPR_8	1	Tetratricopeptide repeat
TPT	1	Triose-phosphate Transporter family
TRANSFERASE	1	Transferase family
TRANSPOSASE_24	1	Plant transposase (Ptta/En/Spm family)
TREHALOSE_PPASE	1	Trehalose-phosphatase
TRIGGER_N	1	Bacterial trigger factor protein (TF)
UBIA	1	UbiA prenyltransferase family
UBIQUITIN	1	Ubiquitin family
UCH	1	Ubiquitin carboxyl-terminal hydrolase
UCR_14KD	1	Ubiquinol-cytochrome C reductase complex 14kD subunit
UTP11	1	Utp11 protein
VARLMGL	1	DUF761-associated sequence motif
WAX2_C	1	WAX2 C-terminal domain
WHEP-TRS	1	WHEP-TRS domain
WRC	1	WRC
WRKY	1	WRKY DNA -binding domain
X8	1	X8 domain
YABBY	1	YABBY protein
YBAB_DNA_BD	1	YbaB/EbfC DNA-binding family
YIP1	1	Yip1 domain
YIPPEE-MIS18	1	Yippee zinc-binding/DNA-binding /Mis18, centromere assembly
YSC84	1	Las17-binding protein actin regulator
ZEIN-BINDING	1	Zein-binding
ZF-A20	1	A20-like zinc finger
ZF-AN1	1	AN1-like Zinc finger
ZF-C3HC4_2	1	Zinc finger, C3HC4 type (RING finger)
ZF-C6H2	1	zf-MYND-like zinc finger, mRNA-binding
ZF-CCCH_2	1	RNA-binding, Nab2-type zinc finger
ZF-MET	1	Zinc-finger of C2H2 type
ZF-NADH-PPASE	1	NADH pyrophosphatase zinc ribbon domain
ZF-NOSIP	1	Zinc-finger of nitric oxide synthase-interacting protein
ZF-RING_2	1	Ring finger domain
ZF-UDP	1	Zinc-binding RING-finger
GRAND TOTAL	439	

Tabla S 7. Genes Región YLD2can.

GENES	NO. DE GENES	DESCRIPCIÓN
P450	14	Cytochrome P450
PKINASE	11	Protein kinase domain
NB-ARC	9	NB-ARC domain
LRR_8	5	Leucine rich repeat
MALECTIN	5	Di-glucose binding within endoplasmic reticulum
RRM_1	5	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
DIMERISATION	4	Dimerisation domain
INHIBITOR_I9	4	Peptidase inhibitor I9
MATE	4	MatE
MYB_DNA-BINDING	4	Myb-like DNA-binding domain
PKINASE_TYR	4	Protein tyrosine kinase
RAS	4	Ras family
TPT	4	Triose-phosphate Transporter family
AMINOTRAN_1_2	3	Aminotransferase class I and II
EF-HAND_7	3	EF-hand domain pair
FBD	3	FBD
FMN_DH	3	FMN-dependent dehydrogenase
GLYCO_HYDRO_1	3	Glycosyl hydrolase family 1
GST_N	3	Glutathione S-transferase, N-terminal domain
HLH	3	Helix-loop-helix DNA-binding domain
METHYLTRANSF_11	3	Methyltransferase domain
NAM	3	No apical meristem (NAM) protein
NB-LRR	3	Late blight resistance protein R1
PEPCASE	3	Phosphoenolpyruvate carboxylase
PMR5N	3	PMR5 N terminal Domain
SUI1	3	Translation initiation factor SUI1
TERPENE_SYNTH_C	3	Terpene synthase family, metal binding domain
UQ_CON	3	Ubiquitin-conjugating enzyme
WD40	3	WD domain, G-beta repeat
X8	3	X8 domain
XET_C	3	Xyloglucan endo-transglycosylase (XET) C-terminus
AA_TRANS	2	Transmembrane amino acid transporter protein
AAA	2	ATPase family associated with various cellular activities (AAA)
ACONITASE_C	2	Aconitase C-terminal domain
ANF_RECEPTOR	2	Receptor family ligand binding region

ATG8	2	Autophagy protein Atg8 ubiquitin like
BET	2	Bromodomain extra-terminal - transcription regulation
C2	2	C2 domain
CELLULOSE_SYNT	2	Cellulose synthase
CHLOROA_B-BIND	2	Chlorophyll A-B binding protein
DNA_POL_A_EXO1	2	3'-5' exonuclease
DUF563	2	Protein of unknown function (DUF563)
FH2	2	Formin Homology 2 Domain
FTCD_N	2	Formiminotransferase domain, N-terminal subdomain
GALACTOSYL_T	2	Galactosyltransferase
GLUCAN_SYNTHASE	2	1,3-beta-glucan synthase component
GLYCO_TRANSF_8	2	Glycosyl transferase family 8
GST_C_2	2	Glutathione S-transferase, C-terminal domain
KINESIN	2	Kinesin motor domain
METHYLTRANSF_2	2	O-methyltransferase
PAN_2	2	PAN-like domain
PC-ESTERASE	2	GDSL/SGNH-like Acyl-Esterase family found in Pmr5 and Cas1p
PP2C	2	Protein phosphatase 2C
PPR	2	PPR repeat
PPR_2	2	PPR repeat family
PYR_REDOX_2	2	Pyridine nucleotide-disulphide oxidoreductase
SOLUTE_TRANS_A	2	Organic solute transporter Ostalpha
TCP	2	TCP family transcription factor
TREHALOSE_PPASE	2	Trehalose-phosphatase
UCH	2	Ubiquitin carboxyl-terminal hydrolase
WRKY	2	WRKY DNA -binding domain
JAS	1	TPL-binding domain in jasmonate signalling
14-3-3	1	14-3-3 protein
2-OXOACID_DH	1	2-oxoacid dehydrogenases acyltransferase (catalytic domain)
3HCDH	1	3-hydroxyacyl-CoA dehydrogenase, C-terminal domain
3HCDH_N	1	3-hydroxyacyl-CoA dehydrogenase, NAD binding domain
ABC1	1	ABC1 family
ABC2_MEMBRANE	1	ABC-2 type transporter
ACYL-COA_DH_M	1	Acyl-CoA dehydrogenase, middle domain
ACYLTRANSFERASE	1	Acyltransferase
ADH_SHORT	1	short chain dehydrogenase
ADH_ZINC_N_2	1	Zinc-binding dehydrogenase
AGENET	1	Agenet domain

ALA_RACEMASE_N	1	Alanine racemase, N-terminal domain
AMIDASE	1	Amidase
ANAPC4_WD40	1	Anaphase-promoting complex subunit 4 WD40 domain
AP2	1	AP2 domain
API5	1	Apoptosis inhibitory protein 5 (API5)
ARFGAP	1	Putative GTPase activating protein for Arf
ATP-SYNT_D	1	ATP synthase subunit D
AVRRPT-CLEAVAGE	1	Cleavage site for pathogenic type III effector avirulence factor Avr
B3	1	B3 DNA binding domain
BAND_7	1	SPFH domain / Band 7 family
BCAS3	1	Breast carcinoma amplified sequence 3
BES1_N	1	BES1/BZR1 plant transcription factor, N-terminal
BHLH-MYC_N	1	bHLH-MYC and R2R3-MYB transcription factors N-terminal
BRANCH	1	Core-2/I-Branching enzyme
BRIX	1	Brix domain
BROMODOMAIN	1	Bromodomain
BT1	1	BT1 family
BYSTIN	1	Bystin
CAF1	1	CAF1 family ribonuclease
CBM_48	1	Carbohydrate-binding module 48 (Isoamylase N-terminal domain)
CDC27	1	DNA polymerase subunit Cdc27
CDC48_2	1	Cell division protein 48 (CDC48), domain 2
CH	1	Calponin homology (CH) domain
CHOLINE_TRANSPO	1	Plasma-membrane choline transporter
CITRATE_SYNT	1	Citrate synthase, C-terminal domain
CNMP_BINDING	1	Cyclic nucleotide-binding domain
COBRA	1	COBRA-like protein
CORA	1	CorA-like Mg2+ transporter protein
CORC_HLYC	1	Transporter associated domain
COX17	1	Cytochrome C oxidase copper chaperone (COX17)
CRS1_YHBY	1	CRS1 / YhbY (CRM) domain
CS	1	CS domain
CTP_TRANSF_3	1	Cytidylyltransferase
CU_BIND_LIKE	1	Plastocyanin-like domain
CU-OXIDASE_3	1	Multicopper oxidase
CYCLIN_C	1	Cyclin, C-terminal domain
CYTB6-F_FE-S	1	Cytochrome B6-F complex Fe-S subunit
DAGK_CAT	1	Diacylglycerol kinase catalytic domain

DDRGK	1	DDRGK domain
DEK_C	1	DEK C terminal domain
DIOX_N	1	non-haem dioxygenase in morphine synthesis N-terminal
DJ-1_PFPI	1	DJ-1/PfpI family
DNAJ_C	1	DnaJ C terminal domain
DND1_DSRM	1	double strand RNA binding domain from DEAD END PROTEIN 1
DSBD	1	Cytochrome C biogenesis protein transmembrane region
DUF1218	1	Protein of unknown function (DUF1218)
DUF1223	1	Protein of unknown function (DUF1223)
DUF1298	1	Protein of unknown function (DUF1298)
DUF1421	1	UBA-like domain (DUF1421)
DUF1771	1	Domain of unknown function (DUF1771)
DUF212	1	Divergent PAP2 family
DUF3082	1	Protein of unknown function (DUF3082)
DUF3537	1	Protein of unknown function (DUF3537)
DUF3700	1	Aluminium induced protein
DUF4094	1	Domain of unknown function (DUF4094)
DUF4217	1	Domain of unknown function (DUF4217)
DUF4408	1	Domain of unknown function (DUF4408)
DUF4414	1	Domain of unknown function (DUF4414)
DUF455	1	Protein of unknown function (DUF455)
DUF572	1	Family of unknown function (DUF572)
DUF588	1	Domain of unknown function (DUF588)
DUF604	1	Protein of unknown function, DUF604
DUF616	1	Protein of unknown function (DUF616)
DUF641	1	Plant protein of unknown function (DUF641)
DUF688	1	Protein of unknown function (DUF688)
DUF92	1	Integral membrane protein DUF92
DYNAMIN_M	1	Dynamin central region
EAP30	1	EAP30/Vps36 family
ECH_2	1	Enoyl-CoA hydratase/isomerase
EF_TS	1	Elongation factor TS
EF-HAND_4	1	Cytoskeletal-regulatory complex EF hand
ELF1	1	Transcription elongation factor Elf1 like
ELONGIN_A	1	RNA polymerase II transcription factor SIII (Elongin) subunit A
EMP24_GP25L	1	emp24/gp25L/p24 family/GOLD
EMP70	1	Endomembrane protein 70
EPIMERASE	1	NAD dependent epimerase/dehydratase family
ERGIC_N	1	Endoplasmic Reticulum-Golgi Intermediate Compartment (ERGIC)

ERO1	1	Endoplasmic Reticulum Oxidoreductin 1 (ERO1)
ETF	1	Electron transfer flavoprotein domain
F-BOX	1	F-box domain
FBA_1	1	F-box associated
FBA_3	1	F-box associated domain
FER4_13	1	4Fe-4S single cluster domain of Ferredoxin I
FHA	1	FHA domain
FKS1_DOM1	1	1,3-beta-glucan synthase subunit FKS1, domain-1
FMN_RED	1	NADPH-dependent FMN reductase
FRIGIDA	1	Frigida-like protein
FTSZ_C	1	FtsZ family, C-terminal domain
GAR1	1	Gar1/Naf1 RNA binding region
GDPD	1	Glycerophosphoryl diester phosphodiesterase family
GH3	1	GH3 auxin-responsive promoter
GLYCO_HYDRO_14	1	Glycosyl hydrolase family 14
GLYCO_HYDRO_16	1	Glycosyl hydrolases family 16
GLYCO_HYDRO_9	1	Glycosyl hydrolase family 9
GLYCO_TRANSF_20	1	Glycosyltransferase family 20
GLYCO_TRANSF_22	1	Alg9-like mannosyltransferase family
GLYCOS_TRANSF_1	1	Glycosyl transferases group 1
GRAS	1	GRAS domain family
GRPE	1	GrpE
GTP_EFTU_D2	1	Elongation factor Tu domain 2
GUB_WAK_BIND	1	Wall-associated receptor kinase galacturonan-binding
HAD_2	1	Haloacid dehalogenase-like hydrolase
HATPASE_C	1	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase
HAUS4	1	HAUS augmin-like complex subunit 4
HGMA	1	homogentisate 1,2-dioxygenase
HIS_BIOSYNTH	1	Histidine biosynthesis protein
HISTONE_H2A_C	1	C-terminus of histone H2A
HYDROLASE	1	haloacid dehalogenase-like hydrolase
IF-2	1	Translation-initiation factor 2
ILVN	1	Acetohydroxy acid isomeroreductase, NADPH-binding domain
INCENP_ARK-BIND	1	Inner centromere protein, ARK binding region
INHIBITOR_I29	1	Cathepsin propeptide inhibitor domain (I29)
INOSITOL_P	1	Inositol monophosphatase family
ION_TRANS	1	Ion transport protein
ION_TRANS_2	1	Ion channel

IPPT	1	IPP transferase
IQ	1	IQ calmodulin-binding motif
LACTAMASE_B	1	Metallo-beta-lactamase superfamily
LANC_LIKE	1	Lanthionine synthetase C-like protein
LBP_BPI_CETP_C	1	LBP / BPI / CETP family, C-terminal domain
LDH_1_C	1	lactate/malate dehydrogenase, alpha/beta C-terminal domain
LEA_4	1	Late embryogenesis abundant protein
LIPASE_3	1	Lipase (class 3)
LIPOCALIN_2	1	Lipocalin-like domain
LPXK	1	Tetraacyldisaccharide-1-P 4'-kinase
LTP_2	1	Probable lipid transfer
LYSINE_DECARBOX	1	Possible lysine decarboxylase
MAD3_BUB1_I	1	Mad3/BUB1 homology region 1
MAP65_ASE1	1	Microtubule associated protein (MAP65/ASE1 family)
MBF1	1	Multiprotein bridging factor 1
MET_GAMMA_LYASE	1	Methionine gamma-lyase
METALLOPHOS	1	Calcineurin-like phosphoesterase
METHYLTR_RSMB-F	1	16S rRNA methyltransferase RsmB/F
METHYLTRANSF_23	1	Methyltransferase domain
MLO	1	Mlo family
MO25	1	Mo25-like
MOTILE_SPERM	1	MSP (Major sperm protein) domain
MPV17_PMP22	1	Mpv17 / PMP22 family
MTERF	1	mTERF
MTN3_SLV	1	Sugar efflux transporter for intercellular exchange
MUTS_V	1	MutS domain V
NAD_BINDING_10	1	NAD(P)H-binding
NADH-U_OX-RDASE	1	NADH-ubiquinone oxidoreductase complex I, 21 kDa subunit
NGP1NT	1	NGP1NT (NUC091) domain
NHL	1	NHL repeat
NIF	1	NLI interacting factor-like phosphatase
NIFU	1	NifU-like domain
NTP_TRANSFERASE	1	Nucleotidyl transferase
NUDIX	1	NUDIX domain
OPA3	1	Optic atrophy 3 protein (OPA3)
OTCACE_N	1	Aspartate/ornithine carbamoyltransferase, carbamoyl-P binding domain
PA	1	PA domain

PALP	1	Pyridoxal-phosphate dependent enzyme
PAM2	1	Ataxin-2 C-terminal region
PAP_ASSOC	1	Cid1 family poly A polymerase
PCO_ADO	1	PCO_ADO
PDT	1	Prephenate dehydratase
PEC_LYASE_C	1	Pectate lyase
PECTINESTERASE	1	Pectinesterase
PEPTIDASE_C13	1	Peptidase C13 family
PFKB	1	pfkB family carbohydrate kinase
PGP_PHOSPHATASE	1	Mitochondrial PGP phosphatase
PHD	1	PHD-finger
PIRIN_C	1	Pirin C-terminal cupin domain
PLUG_TRANSLOCON	1	Plug domain of Sec61p
PNK3P	1	Polynucleotide kinase 3 phosphatase
POLLEN_OLE_E_I	1	Pollen proteins Ole e I like
POLYPRENYL_SYNT	1	Polyprenyl synthetase
POX_MCEL	1	mRNA capping enzyme
PPR_3	1	Pentatricopeptide repeat domain
PPR_LONG	1	Pentacotripeptide-repeat region of PRORP
PPTA	1	Protein prenyltransferase alpha subunit repeat
PRA1	1	PRA1 family protein
PRIBOSYLTRAN_N	1	N-terminal domain of ribose phosphate pyrophosphokinase
PRMA	1	Ribosomal protein L11 methyltransferase (PrmA)
PROTEASOM_PSMB	1	Proteasome non-ATPase 26S subunit
PTERIN_4A	1	Pterin 4 alpha carbinolamine dehydratase
PUR_AC_PHOSPH_N	1	Purple acid Phosphatase, N-terminal domain
RAB3-GTPASE_CAT	1	Rab3 GTPase-activating protein catalytic subunit
RAB3GAP2_N	1	Rab3 GTPase-activating protein regulatory subunit N-terminus
RCC1	1	Regulator of chromosome condensation (RCC1) repeat
RCC1_2	1	Regulator of chromosome condensation (RCC1) repeat
RCD1	1	Cell differentiation family, Rcd1-like
RIBONUC_L-PSP	1	Endoribonuclease L-PSP
RIBOSOMAL_L1	1	Ribosomal protein L1p/L10e family
RIBOSOMAL_L13	1	Ribosomal protein L13
RIBOSOMAL_L2_C	1	Ribosomal Proteins L2, C-terminal domain
RIBOSOMAL_L5_C	1	ribosomal L5P family C-terminus
RIBOSOMAL_S14	1	Ribosomal protein S14p/S29e
RIBOSOMAL_S20P	1	Ribosomal protein S20

RNASE_T	1	Exonuclease
RRM_2	1	RNA recognition motif 2
SAWADEE	1	SAWADEE domain
SBF	1	Sodium Bile acid symporter family
SBP	1	SBP domain
SCAMP	1	SCAMP family
SELR	1	SelR domain
SM-ATX	1	Ataxin 2 SM domain
SNARE	1	SNARE domain
SNF2_N	1	SNF2 family N-terminal domain
SOF1	1	Sof1-like domain
SOUL	1	SOUL heme-binding protein
SPG48	1	AP-5 complex subunit, vesicle trafficking
STR_SYNTH	1	Strictosidine synthase
SUGAR_TR	1	Sugar (and other) transporter
SURP	1	Surp module
SWIB	1	SWIB/MDM2 domain
TAF	1	TATA box binding protein associated factor (TAF)
TAF1_SUBA	1	TAF RNA Polymerase I subunit A
TB2_DP1_HVA22	1	TB2/DP1, HVA22 family
THI4	1	Thi4 family
THIF	1	ThiF family
THIOREDOXIN	1	Thioredoxin
THYLAKOIDFORMAT	1	Thylakoid formation protein
TMEMB_14	1	Transmembrane proteins 14C
TOPO_C_ASSOC	1	C-terminal topoisomerase domain
TPR_1	1	Tetratricopeptide repeat
TPR_16	1	Tetratricopeptide repeat
TPR_2	1	Tetratricopeptide repeat
TPX2_IMPORTIN	1	Cell cycle regulated microtubule associated protein
TRAB	1	TraB family
TRANSKETOLASE_C	1	Transketolase, C-terminal domain
TRAPP	1	Transport protein particle (TRAPP) component
TRAPPC10	1	Trafficking protein particle complex subunit 10, TRAPPC10
TRIGGER_C	1	Bacterial trigger factor protein (TF) C-terminus
TRNA-SYNT_1B	1	tRNA synthetases class I (W and Y)
TUB	1	Tub family
UBD	1	Ubiquitin-binding domain
URO-D	1	Uroporphyrinogen decarboxylase (URO-D)

VPS16_C	1	Vps16, C-terminal region
VTA1	1	Vta1 like
WSD	1	Williams-Beuren syndrome DDT (WSD), D-TOX E motif
ZF-B_BOX	1	B-box zinc finger
ZF-C3HC	1	C3HC zinc finger-like
ZF-C3HC4_3	1	Zinc finger, C3HC4 type (RING finger)
ZF-HC5HC2H	1	PHD-like zinc-binding domain
ZF-MYND	1	MYND finger
ZF-UBP	1	Zn-finger in ubiquitin-hydrolases and other protein
ZIP	1	ZIP Zinc transporter
GRAND TOTAL	442	

Tabla S 8. Genes Región YLD2eug.

GENES	NO. DE GENES	DESCRIPCIÓN
P450	12	Cytochrome P450
PKINASE	12	Protein kinase domain
NB-ARC	8	NB-ARC domain
LRR_8	5	Leucine rich repeat
METHYLTRANSF_2	5	O-methyltransferase
RRM_1	5	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
AA_TRANS	4	Transmembrane amino acid transporter protein
FBD	4	FBD
MALECTIN	4	Di-glucose binding within endoplasmic reticulum
NB-LRR	4	Late blight resistance protein R1
TERPENE_SYNTH_C	4	Terpene synthase family, metal binding domain
AMINOTRAN_1_2	3	Aminotransferase class I and II
FMN_DH	3	FMN-dependent dehydrogenase
GLUCAN_SYNTHASE	3	1,3-beta-glucan synthase component
GLYCO_HYDRO_1	3	Glycosyl hydrolase family 1
GST_N	3	Glutathione S-transferase, N-terminal domain
INHIBITOR_I9	3	Peptidase inhibitor I9
MATE	3	MatE
MYB_DNA-BINDING	3	Myb-like DNA-binding domain

NAM	3	No apical meristem (NAM) protein
PKINASE_TYR	3	Protein tyrosine kinase
PPR	3	PPR repeat
SUI1	3	Translation initiation factor SUI1
TPT	3	Triose-phosphate Transporter family
WD40	3	WD domain, G-beta repeat
AAA	2	ATPase family associated with various cellular activities (AAA)
ACONITASE_C	2	Aconitase C-terminal domain
ANF_RECEPTOR	2	Receptor family ligand binding region
ATG8	2	Autophagy protein Atg8 ubiquitin like
BET	2	Bromodomain extra-terminal - transcription regulation
C2	2	C2 domain
CHLOROA_B-BIND	2	Chlorophyll A-B binding protein
DEK_C	2	DEK C terminal domain
DNA_POL_A_EXO1	2	3'-5' exonuclease
DUF563	2	Protein of unknown function (DUF563)
EF-HAND_7	2	EF-hand domain pair
FH2	2	Formin Homology 2 Domain
GALACTOSYL_T	2	Galactosyltransferase
GH3	2	GH3 auxin-responsive promoter
GLYCO_TRANSF_22	2	Alg9-like mannosyltransferase family
GLYCO_TRANSF_8	2	Glycosyl transferase family 8
GST_C_2	2	Glutathione S-transferase, C-terminal domain
HLH	2	Helix-loop-helix DNA-binding domain
KINESIN	2	Kinesin motor domain
METHYLTRANSF_11	2	Methyltransferase domain
PC-ESTERASE	2	GDSL/SGNH-like Acyl-Esterase family found in Pmr5 and Cas1p
PMR5N	2	PMR5 N terminal Domain
PP2C	2	Protein phosphatase 2C
PPR_2	2	PPR repeat family
PYR_REDOX_2	2	Pyridine nucleotide-disulphide oxidoreductase
RAS	2	Ras family
RIBOSOMAL_L2_C	2	Ribosomal Proteins L2, C-terminal domain
UCH	2	Ubiquitin carboxyl-terminal hydrolase
UQ_CON	2	Ubiquitin-conjugating enzyme
X8	2	X8 domain
XET_C	2	Xyloglucan endo-transglycosylase (XET) C-terminus
JAS	1	TPL-binding domain in jasmonate signalling

2-OXOACID_DH	1	2-oxoacid dehydrogenases acyltransferase (catalytic domain)
3HCDH	1	3-hydroxyacyl-CoA dehydrogenase, C-terminal domain
3HCDH_N	1	3-hydroxyacyl-CoA dehydrogenase, NAD binding domain
ABC1	1	ABC1 family
ABC2_MEMBRANE	1	ABC-2 type transporter
ACOX	1	Acyl-CoA oxidase
ACYLTRANSFERASE	1	Acyltransferase
ADH_SHORT	1	short chain dehydrogenase
ADH_ZINC_N_2	1	Zinc-binding dehydrogenase
AGENET	1	Agenet domain
AMIDASE	1	Amidase
ANK_2	1	Ankyrin repeats (3 copies)
AP2	1	AP2 domain
API5	1	Apoptosis inhibitory protein 5 (API5)
ARFGAP	1	Putative GTPase activating protein for Arf
ATP-SYNT_D	1	ATP synthase subunit D
AVRRPT-CLEAVAGE	1	Cleavage site for pathogenic type III effector avirulence factor Avr
B3	1	B3 DNA binding domain
BAND_7	1	SPFH domain / Band 7 family
BCAS3	1	Breast carcinoma amplified sequence 3
BES1_N	1	BES1/BZR1 plant transcription factor, N-terminal
BHLH-MYC_N	1	bHLH-MYC and R2R3-MYB transcription factors N-terminal
BRANCH	1	Core-2/I-Branching enzyme
BRIX	1	Brix domain
BROMODOMAIN	1	Bromodomain
BT1	1	BT1 family
BYSTIN	1	Bystin
BZIP_2	1	Basic region leucine zipper
CBM_48	1	Carbohydrate-binding module 48 (Isoamylase N-terminal domain)
CDC27	1	DNA polymerase subunit Cdc27
CELLULASE	1	Cellulase (glycosyl hydrolase family 5)
CELLULOSE_SYNT	1	Cellulose synthase
CH	1	Calponin homology (CH) domain
CHOLINE_TRANSPO	1	Plasma-membrane choline transporter
CITRATE_SYNT	1	Citrate synthase, C-terminal domain
CNMP_BINDING	1	Cyclic nucleotide-binding domain
COBRA	1	COBRA-like protein
CORA	1	CorA-like Mg2+ transporter protein

CORC_HLYC	1	Transporter associated domain
COX17	1	Cytochrome C oxidase copper chaperone (COX17)
CRS1_YHBY	1	CRS1 / YhbY (CRM) domain
CS	1	CS domain
CTP_TRANSF_3	1	Cytidylyltransferase
CU_BIND_LIKE	1	Plastocyanin-like domain
CU-OXIDASE_3	1	Multicopper oxidase
CYCLIN_C	1	Cyclin, C-terminal domain
CYTB6-F_FE-S	1	Cytochrome B6-F complex Fe-S subunit
DDRGK	1	DDRGK domain
DIMERISATION	1	Dimerisation domain
DIOX_N	1	non-haem dioxygenase in morphine synthesis N-terminal
DJ-1_PFPI	1	DJ-1/PfpI family
DNAJ	1	DnaJ domain
DND1_DSRM	1	double strand RNA binding domain from DEAD END PROTEIN 1
DSBD	1	Cytochrome C biogenesis protein transmembrane region
DUF1218	1	Protein of unknown function (DUF1218)
DUF1223	1	Protein of unknown function (DUF1223)
DUF1298	1	Protein of unknown function (DUF1298)
DUF1421	1	UBA-like domain (DUF1421)
DUF1771	1	Domain of unknown function (DUF1771)
DUF212	1	Divergent PAP2 family
DUF2781	1	Protein of unknown function (DUF2781)
DUF3082	1	Protein of unknown function (DUF3082)
DUF3403	1	Domain of unknown function (DUF3403)
DUF3537	1	Protein of unknown function (DUF3537)
DUF3700	1	Aluminium induced protein
DUF4094	1	Domain of unknown function (DUF4094)
DUF4408	1	Domain of unknown function (DUF4408)
DUF4414	1	Domain of unknown function (DUF4414)
DUF572	1	Family of unknown function (DUF572)
DUF588	1	Domain of unknown function (DUF588)
DUF604	1	Protein of unknown function, DUF604
DUF616	1	Protein of unknown function (DUF616)
DUF641	1	Plant protein of unknown function (DUF641)
DUF92	1	Integral membrane protein DUF92
EAP30	1	EAP30/Vps36 family
ECH_2	1	Enoyl-CoA hydratase/isomerase
EF_TS	1	Elongation factor TS
EF-HAND_4	1	Cytoskeletal-regulatory complex EF hand

ELF1	1	Transcription elongation factor Elf1 like
ELONGIN_A	1	RNA polymerase II transcription factor SIII (Elongin) subunit A
EMP24_GP25L	1	emp24/gp25L/p24 family/GOLD
EMP70	1	Endomembrane protein 70
EPIMERASE	1	NAD dependent epimerase/dehydratase family
ERGIC_N	1	Endoplasmic Reticulum-Golgi Intermediate Compartment (ERGIC)
ERO1	1	Endoplasmic Reticulum Oxidoreductin 1 (ERO1)
ETF	1	Electron transfer flavoprotein domain
F-BOX	1	F-box domain
FBA_1	1	F-box associated
FHA	1	FHA domain
FKS1_DOM1	1	1,3-beta-glucan synthase subunit FKS1, domain-1
FMN_RED	1	NADPH-dependent FMN reductase
FRIGIDA	1	Frigida-like protein
FTCD_N	1	Formiminotransferase domain, N-terminal subdomain
FTSZ_C	1	FtsZ family, C-terminal domain
GLYCO_HYDRO_14	1	Glycosyl hydrolase family 14
GLYCO_HYDRO_16	1	Glycosyl hydrolases family 16
GLYCO_HYDRO_9	1	Glycosyl hydrolase family 9
GLYCOS_TRANSF_1	1	Glycosyl transferases group 1
GRAS	1	GRAS domain family
GRPE	1	GrpE
GTP_EFTU_D2	1	Elongation factor Tu domain 2
GUB_WAK_BIND	1	Wall-associated receptor kinase galacturonan-binding
HATPASE_C	1	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase
HAUS4	1	HAUS augmin-like complex subunit 4
HELICASE_C	1	Helicase conserved C-terminal domain
HGMA	1	homogentisate 1,2-dioxygenase
HIS_BIOSYNTH	1	Histidine biosynthesis protein
HISTONE_H2A_C	1	C-terminus of histone H2A
HYDROLASE	1	haloacid dehalogenase-like hydrolase
IF-2	1	Translation-initiation factor 2
ILVN	1	Acetohydroxy acid isomeroreductase, NADPH-binding domain
INCENP_ARK-BIND	1	Inner centromere protein, ARK binding region
INHIBITOR_I29	1	Cathepsin propeptide inhibitor domain (I29)
INOSITOL_P	1	Inositol monophosphatase family
ION_TRANS_2	1	Ion channel
IQ	1	IQ calmodulin-binding motif

LACTAMASE_B	1	Metallo-beta-lactamase superfamily
LBP_BPI_CETP_C	1	LBP / BPI / CETP family, C-terminal domain
LDH_1_C	1	lactate/malate dehydrogenase, alpha/beta C-terminal domain
LIAS_N	1	N-terminal domain of lipoyl synthase of Radical_SAM family
LIPASE_3	1	Lipase (class 3)
LIPOCALIN_2	1	Lipocalin-like domain
LPXK	1	Tetraacyldisaccharide-1-P 4'-kinase
LRRNT_2	1	Leucine rich repeat N-terminal domain
LYSINE_DECARBOX	1	Possible lysine decarboxylase
MAP65_ASE1	1	Microtubule associated protein (MAP65/ASE1 family)
MBF1	1	Multiprotein bridging factor 1
MET_GAMMA_LYASE	1	Methionine gamma-lyase
METALLOPHOS	1	Calcineurin-like phosphoesterase
METHYLTRANSF_23	1	Methyltransferase domain
MLO	1	Mlo family
MO25	1	Mo25-like
MOTILE_SPERM	1	MSP (Major sperm protein) domain
MPV17_PMP22	1	Mpv17 / PMP22 family
MTERF	1	mTERF
MTN3_SLV	1	Sugar efflux transporter for intercellular exchange
MUTS_V	1	MutS domain V
NAD_BINDING_10	1	NAD(P)H-binding
NADH-U_OX-RDASE	1	NADH-ubiquinone oxidoreductase complex I, 21 kDa subunit
NGP1NT	1	NGP1NT (NUC091) domain
NHL	1	NHL repeat
NIF	1	NLI interacting factor-like phosphatase
NIFU	1	NifU-like domain
NTP_TRANSFERASE	1	Nucleotidyl transferase
OTCACE_N	1	Aspartate/ornithine carbamoyltransferase, carbamoyl-P binding domain
PA	1	PA domain
PALP	1	Pyridoxal-phosphate dependent enzyme
PAM2	1	Ataxin-2 C-terminal region
PAN_2	1	PAN-like domain
PAP_ASSOC	1	Cid1 family poly A polymerase
PDT	1	Prephenate dehydratase
PEC_LYASE_C	1	Pectate lyase
PECTINESTERASE	1	Pectinesterase
PEPCASE	1	Phosphoenolpyruvate carboxylase

PEPTIDASE_C13	1	Peptidase C13 family
PFKB	1	pfkB family carbohydrate kinase
PGP_PHOSPHATASE	1	Mitochondrial PGP phosphatase
PHD	1	PHD-finger
PIRIN_C	1	Pirin C-terminal cupin domain
PLUG_TRANSLOCON	1	Plug domain of Sec61p
PNK3P	1	Polynucleotide kinase 3 phosphatase
POLLEN_OLE_E_I	1	Pollen proteins Ole e I like
POLYPRENYL_SYNT	1	Polyprenyl synthetase
PPR_3	1	Pentatricopeptide repeat domain
PPR_LONG	1	Pentacotripeptide-repeat region of PRORP
PPTA	1	Protein prenyltransferase alpha subunit repeat
PRA1	1	PRA1 family protein
PRIBOSYLTRAN_N	1	N-terminal domain of ribose phosphate pyrophosphokinase
PRMA	1	Ribosomal protein L11 methyltransferase (PrmA)
PROTEASOM_PSMB	1	Proteasome non-ATPase 26S subunit
PUR_AC_PHOSPH_N	1	Purple acid Phosphatase, N-terminal domain
RAB3GAP2_N	1	Rab3 GTPase-activating protein regulatory subunit N-terminus
RCC1	1	Regulator of chromosome condensation (RCC1) repeat
RCC1_2	1	Regulator of chromosome condensation (RCC1) repeat
RCD1	1	Cell differentiation family, Rcd1-like
RIBONUC_L-PSP	1	Endoribonuclease L-PSP
RIBOSOMAL_L13	1	Ribosomal protein L13
RIBOSOMAL_L5_C	1	ribosomal L5P family C-terminus
RIBOSOMAL_S20P	1	Ribosomal protein S20
RNASE_T	1	Exonuclease
RRM_2	1	RNA recognition motif 2
SAWADEE	1	SAWADEE domain
SBF	1	Sodium Bile acid symporter family
SBP	1	SBP domain
SELR	1	SelR domain
SNARE	1	SNARE domain
SNF2_N	1	SNF2 family N-terminal domain
SOF1	1	Sof1-like domain
SOLUTE_TRANS_A	1	Organic solute transporter Ostalpha
SOUL	1	SOUL heme-binding protein
SPG48	1	AP-5 complex subunit, vesicle trafficking
STR_SYNTH	1	Strictosidine synthase

SUGAR_TR	1	Sugar (and other) transporter
SURP	1	Surp module
SWIB	1	SWIB/MDM2 domain
TAF1_SUBA	1	TAF RNA Polymerase I subunit A
TAF6_C	1	TAF6 C-terminal HEAT repeat domain
TCP	1	TCP family transcription factor
THI4	1	Thi4 family
THIF	1	ThiF family
THIOREDOXIN	1	Thioredoxin
THYLAKOIDFORMAT	1	Thylakoid formation protein
TOPO_C_ASSOC	1	C-terminal topoisomerase domain
TPR_1	1	Tetratricopeptide repeat
TPR_16	1	Tetratricopeptide repeat
TPR_2	1	Tetratricopeptide repeat
TRAB	1	TraB family
TRANSKETOLASE_C	1	Transketolase, C-terminal domain
TRAPP	1	Transport protein particle (TRAPP) component
TRAPPC10	1	Trafficking protein particle complex subunit 10, TRAPPC10
TREHALOSE_PPASE	1	Trehalose-phosphatase
TRIGGER_C	1	Bacterial trigger factor protein (TF) C-terminus
TRNA-SYNT_1B	1	tRNA synthetases class I (W and Y)
TUB	1	Tub family
UBD	1	Ubiquitin-binding domain
URO-D	1	Uroporphyrinogen decarboxylase (URO-D)
VPS16_C	1	Vps16, C-terminal region
VTA1	1	Vta1 like
WHIM1	1	WSTF, HB1, Itc1p, MBD9 motif 1
WRKY	1	WRKY DNA -binding domain
ZF-C3HC	1	C3HC zinc finger-like
ZF-C3HC4_3	1	Zinc finger, C3HC4 type (RING finger)
ZF-HC5HC2H	1	PHD-like zinc-binding domain
ZF-MYND	1	MYND finger
ZF-RVT	1	zinc-binding in reverse transcriptase
ZF-UBP	1	Zn-finger in ubiquitin-hydrolases and other protein
ZIP	1	ZIP Zinc transporter
GRAND TOTAL	399	

Anexo IV. Anotación de los genes compartidos por cada una de las regiones asociados dentro de la región flanqueada por marcadores.

NOMBRE DE LA REGIÓN	TOTAL	DESCRIPCIÓN
BS4Ea PH1Ca PH1Ea PH5Ca PH5Ea YLD2Ca YLD2Ea	1	Pkinase
PH1Ca PH1Ea PH4Ea PH5Ca PH5Ea YLD2Ca YLD2Ea	2	NB-ARC
		Pkinase_Tyr
PH1Ca PH1Ea PH5Ca PH5Ea YLD2Ca YLD2Ea	1	PPR_2
PH1Ca PH1Ea PH5Ea YLD2Ca YLD2Ea	2	p450
		Ras
PH1Ea PH5Ca PH5Ea YLD2Ca YLD2Ea	1	HLH
PH4Ea PH5Ca PH5Ea YLD2Ca YLD2Ea	1	LRR_8
PH1Ca PH1Ea PH4Ea PH5Ea	1	RRM_1
PH1Ca PH1Ea PH5Ca PH5Ea	6	ENT
		LEA_1
		Transferase
		Metallophos
		DYW_deaminase
		Tryp_alpha_amyl
PH1Ca PH1Ea YLD2Ca YLD2Ea	4	GST_N
		FKS1_dom1
		ECH_2
		PPR_3
PH4Ea PH5Ea YLD2Ca YLD2Ea	1	Myb_DNA-binding
PH5Ca PH5Ea YLD2Ca YLD2Ea	6	MatE
		DIOX_N
		EMP70
		UCH
		Dimerisation
		PC-Esterase
PH1Ca PH1Ea PH4Ea	1	RWD
PH1Ca PH1Ea PH5Ca	1	MORN
PH1Ca PH1Ea PH5Ea	2	O-FucT
		zf-CCCH_2
PH1Ca PH1Ea YLD2Ca	2	DnaJ_C
		HAD_2
PH1Ca YLD2Ca YLD2Ea	1	Citrate_synt
PH1Ea PH5Ca PH5Ea	4	HD-ZIP_N
		AA_permease_2
		MIF4G
		zf-UDP

PH4Ea PH5Ca PH5Ea	1	WD40
PH4Ea YLD2Ca YLD2Ea	2	Trehalose_PPase
		XET_C
PH5Ca PH5Ea YLD2Ca	1	WRKY
PH5Ca PH5Ea YLD2Ea	1	DnaJ
PH5Ca YLD2Ca YLD2Ea	4	Pyr_redox_2
		FH2
		PP2C
		TPT
PH5Ea YLD2Ca YLD2Ea	1	SBP
BS4Ea PH4Ea	1	Hydrolase_4
PH1Ca PH1Ea	38	ELMO_CED12
		Mito_carr
		Ribosomal_S7e
		LIN37
		Glycohydro_20b2
		Glyco_transf_22
		Ribosomal_L3
		HALZ
		Glyco_hydro_1
		ILVD_EDD
		Sec23_BS
		OMPdecase
		Plant_NMP1
		UbiA
		Propeptide_C1
		Ribosomal_L37ae
		N_Asn_amidohyd
		DUF1005
		Cu_bind_like
		SCAI
		Neprosin_AP
		tRNA_synt_2f
		Cyclin
		Ubie_methyltran
		Vps62
		PP-binding
		PTR2
		NDK
		Alg6_Alg8
		Glyco_hydro_20
		LSM
		CK_II_beta

		Stm1_N
		ATP-synt_DE_N
		zf-LSD1
		Mob_synth_C
		Cnd1_N
		Ufm1
PH1Ea PH4Ea	1	ABC_membrane
PH1Ea PH5Ca	2	MMR_HSR1
		zf-CCCH
PH4Ea PH5Ca	1	Exo_endo_phos
PH4Ea PH5Ea	2	Myb_DNA-bind_3
		ABC_tran
PH5Ca PH5Ea	52	SRP-alpha_N
		Myb_DNA-bind_6
		Peptidase_M17_N
		DUF3336
		adh_short_C2
		Adenine_glyco
		eIF3_N
		Cytokin-bind
		Zein-binding
		NEMP
		Actin
		Tmemb_14
		FAD_binding_3
		Yippee-Mis18
		PRP4
		RNase_PH_C
		Abhydrolase_1
		Peptidase_S28
		Spc97_Spc98
		CAF1
		COX6B
		Methyltransf_29
		Carb_anhydrase
		Nudix_N_2
		Longin
		Gpi1
		zf-NOSIP
		NAF
		Esterase
		DUF3474
		peroxidase

		ubiquitin
		DAO
		Cation_efflux
		DNA_pol3_delta2
		Alg14
		PCRF
		FA_desaturase
		eIF-5a
		Choline_transpo
		Aldedh
		Mic1
		Glyco_hydro_17
		APO_RNA-bind
		Ribosomal_L21e
		NAD_Gly3P_dh_C
		YbaB_DNA_bd
		Helicase_C
		Sina
		zf-C6H2
		QWRF
		BTB_2
PH5Ea YLD2Ca	1	PCO_ADO
YLD2Ca YLD2Ea	86	SBF
		DUF572
		FMN_dh
		tRNA-synt_1b
		PGP_phosphatase
		GST_C_2
		CDC27
		Pribosyltran_N
		MtN3_slv
		RRM_2
		SOUL
		PfkB
		Terpene_synth_C
		DUF1421
		PAM2
		Vps16_C
		Chloroa_b-bind
		Bystin
		Plug_translocon
		Glyco_hydro_14
		Ion_trans_2

		Elongin_A
		zf-MYND
		HATPase_c
		Lactamase_B
		ThylakoidFormat
		PMR5N
		NTP_transferase
		SelR
		Sof1
		PA
		DUF588
		Methyltransf_11
		Band_7
		Aminotran_1_2
		TPR_1
		zf-C3HC
		IF-2
		FTCD_N
		GTP_EFTU_D2
		ANF_receptor
		DUF92
		DUF604
		Transketolase_C
		Inhibitor_I9
		Ribosomal_L13
		FBA_1
		NifU
		SNARE
		Peptidase_C13
		ERO1
		Methyltransf_2
		BES1_N
		NHL
		Amidase
		Glucan_synthase
		Epimerase
		EF_TS
		BET
		Glyco_hydro_16
		PPR
		DUF1298
		X8
		UQ_con

		UBD
		CorA
		CS
		GUB_WAK_bind
		TraB
		Cyclin_C
		C2
		polyprenyl_synt
		Aa_trans
		Malectin
		Lipase_3
		zf-HC5HC2H
		Methyltransf_23
		DUF641
		RCC1_2
		SUI1
		Ribosomal_L5_C
		CytB6-F_Fe-S
		F-box
		Glyco_transf_8
		ATP-synt_D
		IlvN
BS4Ea	26	Sawadee
		Yif1
		Mekhla
		Fad_binding_4
		Pfk
		Nucleopor_nup85
		Tocopherol_cycl
		Duf810
		Gat
		Gata
		Malic_m
		Yippee-mis18
		Trna-synt_1b
		Hydrolase_3
		Pkinase_tyr
		Aminotran_4
		Enod93
		Peptidase_s10
		Far1
		Lrr_6
		Pnp_udp_1

		Uq_con
		Rpe65
		Profilin
		Ech_1
		Rcc1
PH1Ca	11	OHCU_decarbox
		zf-CCCH_3
		G-patch
		MA3
		TP_methylase
		Sigma70_r2
		Cys_Met_Meta_PP
		Plus-3
		zf-met
		UDPGP
		AMP-binding_C
PH1Ea	27	DUF498
		DDE_Tnp_4
		CBFB_NFYA
		eIF2A
		UPF0565
		Palm_thioest
		Telomerase_RBD
		LCM
		DUF632
		RcbX
		tRNA-synt_2d
		ADH_zinc_N
		Fra10Ac1
		NOG1
		Str_synth
		Wax2_C
		Pribosyl_synth
		F-box-like
		Fer4_7
		Ribosomal_L18p
		KH_1
		AATF-Che1
		Malic_M
		Cpn10
		FMN_red
		GED
		Sigma70_r4

PH4Ea	49	BAH
		LSM14
		RNA_pol_Rpc34
		RNA_pol_Rpb2_5
		NAD_binding_10
		RPE65
		C1_2
		MFMR_assoc
		DUF5110
		PTEN_C2
		EamA
		Cellulose_synt
		JmjN
		PIG-X
		Nucleotid_trans
		Iso_dh
		RST
		COG6
		Adaptin_N
		DUF1336
		RSN1_TM
		Glyco_hydro_31
		Syntaxin_2
		DIM1
		B3
		GWT1
		ACBP
		PPTA
		ThiF
		DAD
		PAE
		CobW_C
		GPAT_N
		Myb_CC_LHEQLE
		Ribosomal_L13e
		PAP_fibrillin
		zf-CCHC_6
		Transglut_core2
		Med26
		Ribosomal_L35p
		2OG-FeII_Oxy_2
		Glyco_hydro_19
		Peptidase_M49

		START
		PHM7_cyt
		AOX
		DUF1664
		DUF647
		DUF1350
PH5Ca	33	Peptidase_M16_M
		CaATP_NAI
		Nic96
		NIF
		DBINO
		RecA
		PUNUT
		TF_Zn_Ribbon
		AA_permease_C
		CH
		Nt_Gln_amidase
		NUC173
		Aha1_N
		NST1
		ASL_C
		NAD_binding_1
		NmrA
		RVT_3
		2OG-FeII_Oxy
		AAA
		DUF563
		ABC2_membrane_3
		DUF3479
		DNA_pol_E_B
		Coq4
		LRR_6
		WW
		SUZ
		HSF_DNA-bind
		Acyltransferase
		EF-hand_6
		Ldh_1_N
		BRF1
PH5Ea	41	Exostosin
		DUF4187
		FAD_binding_6
		Nse4_C

		Methyltransf_10
		GDPD
		FPL
		Peptidase_S10
		Kinesin
		Utp11
		Trigger_N
		WRC
		DUF2048
		Peptidase_S41
		Response_reg
		AHSA1
		LRR_1
		EF-hand_7
		OTU
		PRT_C
		SAP
		PPR_1
		zf-RanBP
		SHMT
		Transposase_24
		zf-C3HC4_2
		Glyco_hydro_18
		Fip1
		EF-hand_5
		Tim17
		DUF773
		zf-A20
		Cation_ATPase
		HRDC
		zf-AN1
		Glycos_transf_4
		Acetyltransf_10
		mTERF
		CaM_binding
		Reticulon
		PDCD7
YLD2Ca	27	PAN_2
		zf-B_box
		Fer4_13
		DUF688
		Solute_trans_a
		TCP

		DUF455
		Mad3_BUB1_I
		Gar1
		Ribosomal_S14
		Ribosomal_L1
		TPX2_importin
		DAGK_cat
		Pterin_4a
		Glyco_transf_20
		Dynamin_M
		LEA_4
		ANAPC4_WD40
		FBA_3
		Acyl-CoA_dh_M
		PEPcase
		Ion_trans
		WSD
		SCAMP
		14-3-3
		TB2_DP1_HVA22
		OPA3
YLD2Ea	14	NB-LRR
		Cellulase
		LIAS_N
		DUF2781
		zf-RVT
		DUF3403
		Ribosomal_L2_C
		Ank_2
		LRRNT_2
		ACOX
		bZIP_2
		GH3
		FBD
		WHIM1